Elementary Applied Analysis

Stephen L. Hobbs
Code 52390, SPAWAR Systems Center, San Diego
619-553-2018
steve.hobbs@navy.mil

January 21, 2015

Contents

0.1 List of Key Algorithms and Formulas

- Gram-Schmidt orthogonalization (in finite or inf dimensions)
- compute e^{At}
- solve an ODE with power series
- variation of parameters formula
- successive approximations to solve ODEs
- geometric series for $(I+B)^{-1}$ when B small
- normal equations, in finite or inf dim'al Hilbert spaces; formula for projection operator
- Fourier series coefficients for any orthonormal set
- separation of variables for PDEs
- finite element (or Galarkin) approximations for elliptic PDEs
- eigen-function expansions for solutions of elliptic, parabolic, and hyperbolic PDEs
- first variation and Euler-Lagrange equations
- various applications and formulas for least squares problems (normal equations)
- conditional expectation $E(x|\mathscr{F}_0)$ approximation for a finite sub-sigma field
- compute the Fourier transform of any box-car function
- amplitude modulation
- ideal band-pass filter
- simplest formulas for TDMA, FDMA, and CDMA
- approximate low-pass filter with circuits (polynomials)
- Shannon's sampling formula
- use the Fourier transform to give formula for solution of linear constant coefficient DEs

0.2 List of Key Theorems and Proofs

- Weierstrass M-test
- e^{At} solves $\dot{x}(t) = Ax(t)$
- power series radius of convergence solving ODEs
- convergence of Picard's successive approximations
- a contraction mapping on a complete metric space has a unique fixed point (know proof)
- a projection onto a closed subspace in a Hilbert space always exists; for finite dimensional subspaces it is given by the normal equations (know proof of finite dimensional case)
- Parseval's equality (know formal proof)
- Riesz representation theorem in Hilbert space (know constructive algorithm)
- Raleigh quotient for eigen-values and eigen-vectors of symmetric matrices
- self-adjoint operators have real eigen-values and orthonormal eigen-vectors (know proof)
- existence of solutions of strictly coercive, self-adjoint, elliptic boundary value problems on domains of any shape (know proof)
- formulas for parabolic and hyperbolic PDEs when given the complete orthonormal eigen-functions of the elliptic operator (know formal proof)
- Fourier transform is isometry of $L^2(\mathbb{R})$ onto $L^2(\mathbb{R})$ (know theorem)
- Shannon sampling theorem (know theorem)

0.3 List of Key Applications

A large amount of technology today is based on mechanical or electrical engineering. Mechanical engineering is very old (people have been building houses, wagons, and ships for hundreds of years) but it continues to be an important part of our world. Mechanical engineers are concerned with the strength of materials and the forces acting on them (whether your house stands in a storm or your iPad doesn't break when you drop it), the heating and cooling of materials (whether your car engine over heats or there is enough steam pressure to drive the turbine that makes your electricity), the flow of fluids (how aerodynamic your car or the plane you fly in is, or whether the fuel of that plane gets pumped smoothly into the jet engines, even if the plane under-goes high-g maneuvers). In the next few decades, or centuries, mechanical engineers will make buildings increasingly resistant to earthquakes, improve the generation of energy, and determine if a functioning magnetically levitating train can be built. Most of the math used in mechanical engineering is related to ordinary or partial differential equations. Extensive use is made of finite element packages.

Electrical engineering is a huge part of our world today, with its emphasis on communications and information. Electrical engineers design and analyze communication systems (telephone, radio, TV, cell phones, computer architectures and machine-machine communications, including the internet). They try to get as much 'information' as possible from device A to device B (with or without wires) using the least energy and bandwidth possible, and with as few errors as possible. Much of the math used in electrical engineering is related to the Fourier transform, least squares estimation, and stochastic processes (because there is noise in almost all 'signals'). Extensive use is made of the fast Fourier transform (FFT).

Information processing is a growing technology with many applications in business and government. It has a certain similarity to the signal processing of electronic signals treated by electrical engineers, but the diverse nature of the 'signals' is much greater. Extensions of classical statistical methods, many using advanced Hilberts space methods, are applied to much larger data sets than ever imagined 20 years ago.

- Newtonian particle mechanics (projectiles, aircraft, satellites, spacecraft): $\mathbf{f} = m\ddot{\mathbf{x}}$, $\mathbf{f} = \mathbf{f}(t, \mathbf{x}, \dot{\mathbf{x}})$
- Newtonian continuum mechanics (strength of materials: cars, trains, aircraft, refrigerators, washing machines, buildings, sky-scrapers, bridges, iPad and cell phone cases, acoustics): $\Delta u(x) = f(x)$, $u_{tt}(t,x) - \Delta u(t,x) = f(t,x)$
- force field equations (gravitation and electromagnetic; aircraft, spacecraft, electric wires, optical fibers, sensors, radio and wireless comm's, radar and infer-red sensors): $\nabla \times E = -B$, $\nabla \times B = J$
- diffusion of heat, chemicals, populations (engine cooling, chemical reactions, chemical engineering and manufacturing): $u_t(t,x) - \Delta u(t,x) = f(t,x)$
- Lagrangian mechanics: $\int_a^b T(\dot{\mathbf{x}}) - V(\mathbf{x})\, dt = \int_a^b \left[\frac{1}{2}\sum m_i(\dot{x}_i^2 + \dot{y}_i^2 + \dot{z}_i^2) - c\sum\sum_{i\neq j} \frac{1}{r_{ij}}\right] dt$
- regularization in statistics and decision theory (business and military decisions, information processing, machine learning, Watson on Jeopardy): $\mathscr{F}(f) = \frac{1}{n}\sum_{i=1}^n V(y_i, f(x_i)) + \lambda \|f\|_K$
- business, economics (forecasting, demand, supply, production, revenue)
- matched filter: $\langle x, m\rangle = \sum_1^n x_i m_i$
- signal processing (transmission of signals and data of all types, both wired and wireless): $f(t) = \int_a^b e^{i\omega t}\hat{f}(\omega)\, d\omega$
- signal and image compression (land-line phone voice): $f(t) \approx g(t) = a_1 h_1(t) + a_2 h_2(t) + a_3 h_3(t)$
- TDMA, FDMA, CDMA (radio, TV, remote controls, cordless phones, GPS, cell phones)

0.4 References

- Arfken, G.; *Mathematical Methods for Physicists*, 3rd ed., Academic Press, 1985.
- Courant, R. and D. Hilbert; *Methods of Mathematical Physics*, v. 1, Wiley-Interscience, 1937.
- Dettman, J. W.; *Mathematical Methods in Physics and Engineering*, (McGraw-Hill, 1962), Dover, 1988.
- Evans; *Partial Differential Equations*, Am Math Soc, 1998.
- Folland, G. B.; *Introduction to Partial Differential Equations*, 2nd ed, Princeton, 1995.
- Fox; *Introduction to the Calculus of Variations*, Dover, 2010.
- Friedman, B; *Principles and Techniques of Applied Mathematics*, Dover, 1990.
- Gelfand, I. M. and S. V. Fomin; *Calculus of Variations*, Dover, 1963.
- Gockenbach, M. S.; *Partial Differential Equations, Analytical and Numerical Methods*, SIAM, 2011.
- Haberman, R.; *Applied Partial Differential Equations: with Fourier Series and Boundary Value Problems*, publ TBD, yr TBD.
- Hubbard, J. and B. B. Hubbard; *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach*, Matrix Editions, 2007.
- Kreyszig, E.; *Introductory Functional Analysis with Applications*, Wiley, 1989.
- Kreyszig, E.; *Advanced Engineering Mathematics*, Wiley, 10ed, 2010.
- Marsden, J. E. and M. J. Hoffman; *Elementary Classical Analysis*, 2nd ed., Freeman, 1993.
- Marsden, J. E.; *Basic Complex Analysis*, Freeman, 1973.
- Moon, T. K. and W. C. Stirling; *Mathematical Methods and Algorithms for Signal Processing*, Prentice Hall, 2000.
- Naylor, A. W. and G. R. Sell; *Linear Operator Theory in Engineering and Science*, Springer, 1982.
- Neter, J., Kutner, Nachtsheim, Wasserman; *Applied Linear Statistical Models*, McGraw Hill, 4ed, 1996.
- Oppenheim, A. V. and A. S. Willsky; *Signals and Systems*, Prentice Hall, 1983.
- Rudin, W.; *Principles of Mathematical Analysis*, 3rd ed., McGraw Hill, 1976.
- Rudin, W.; *Real and Complex Analysis*, 2nd ed., McGraw Hill, 1974.
- Samuelson, W. F. and S. G. Marks; *Managerial Economics*, Wiley, 7ed, 2012.
- Smith, D. R.; *Variational Methods in Optimization*, Prentice Hall (1974), Dover, 1998.
- Stakgold, I.; *Green's Functions and Boundary Value Problems*, Wiley-Interscience, 1979.

0.5 Notation

- sets: compliment A^c, relative compliment $A\backslash B$, closure cl(A) or $\overline{A}$, interior int(A) or A°, boundary ∂A ($=\overline{A}\backslash A^\circ$)
- $A \times B = \{(a,b)\ ;\ a \in A \text{ and } b \in B\}$,
- function $f : X \to Y$, X = domain, Y = range
- continuous functions $C(X;Y)$ from metric space X to metric space Y; a vector space if Y is a vector space
- $\mathbb{N} = \{1,2,...\}$, $\mathbb{N}_0 = \{0,1,2,...\}$, $\mathbb{Z} = \{0,\pm 1,\pm 2,...\}$, $\mathbb{Q}$ = rationals, $\mathbb{R}$ = reals, $\mathbb{C}$ = complex numbers, R^n real (usu. column) n-vectors, $\mathbb{C}^n$ complex (usu. column) n-vectors, $\mathbb{R}^{m\times n}$ real $m \times n$ matrices, $\mathbb{C}^{m\times n}$ complex $m \times n$ matrices
- indicator function $1_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \in A^c \end{cases}$
- real and imaginary parts of complex number (or matrix), $x = \Re z$ and $y = \Im z$ when $z = x + iy$; we may change this notation to Re....
- complex conjugate or complex number (or matrix) $\bar{z}$; transpose of linear transformation T'; conjugate transpose T^*
- $T = [t_{ij}]$ when $t_{i,j}$ is the ij-th component of the matrix T, and perhaps $T = [(t_{i,j})_{i=1,...,m;j=1,...,n}]$ when T is $m \times n$
- Kronecker delta $\delta_{ij} = 1$ when $i = j$, and $\delta_{ij} = 0$ when $i \neq j$.
- dim(X) dimension of the vector space X
- span$\{x_1,...,x_n\}$ span of the vectors $x_1,...,x_n$
- $\mathscr{R}(T)$ range of the linear transformation T
- $\mathscr{N}(T)$ null space (or kernel) of the linear transformation T
- $\Omega \subset \mathbb{R}^d$ open subset (usually tacitly assumed non-empty), domain of space of functions
- Hilbert space inner product $\langle u,v\rangle$ (conjugate linear in u and linear in v) if u and v elements of a Hilbert space; OR duality pairing (linear in both u and v) if u and v elements of a Banach space that is not a Hilbert spaces; some authors use $\langle u|v\rangle$ for one, to distinguish the two -probably a good idea-
- $L^p(\Omega)$
- $H^k(\Omega)$ with inner product $\langle u,v\rangle_k$
- multi-index notation:

$$\partial_j \;=\; \frac{\partial}{\partial x_j}\,, \qquad \partial^\alpha \;=\; \partial_1^{\alpha_1}\cdots\partial_d^{\alpha_d}\,, \qquad x^\alpha \;=\; x_1^{\alpha_1}x_2^{\alpha_2}...x_d^{\alpha_d}\,, \qquad \text{and} \qquad |\alpha| = \alpha_1 + \cdots + \alpha_d$$

when local coordinates $x = (x_1,...,x_d)$ are understood and $\alpha = (\alpha_1,...,\alpha_d)$ is a tuple whose components are non-negative integers.

For example, if $d = 3$ and $\alpha = (1,0,2)$ then $|\alpha| = 3$ and

$$\partial^\alpha \;=\; \partial_1\partial_3^2 \;=\; \frac{\partial^3}{\partial x_1 \partial x_3^2} \qquad \text{and} \qquad x^\alpha \;=\; x_1x_3^2\,.$$

Also, expressions like $\partial_{ij} = \partial_i\partial_j$ and $\partial_{ijk} = \partial_i\partial_j\partial_k$ will be used.

- if $k \in \mathbb{N}_0$, $C^k(\Omega) = C^k(\Omega; \mathbb{K})$ is the vector space of scalar-valued functions on an open set $\Omega \subset \mathbb{R}^d$ which, together with all partial derivatives up through order k, are continuous on Ω. (Usually $C^0(\Omega)$ is denoted $C(\Omega)$.) We also refer to $C^\infty(\Omega) = \cap_{k=0}^\infty C^k(\Omega)$, the vector space of continuous functions which are infinitely differentiable on Ω.[1]
- $\mathrm{spt}(f)$ support of the function f (= closure of the set $\{x\ ;\ f(x) \neq 0\}$)
- if $k \in \mathbb{N}_0$ or $k = \infty$, $C_0^k(\Omega) = C^k$ functions with compact support in Ω; 'test functions'

[1] The vector spaces $C^k(\Omega)$ are not Banach spaces. The natural topology on each of these spaces is given by requiring uniform convergence on compact subsets of the open set Ω. Each of these spaces is a complete metric space, but the metric cannot be given by a norm.

0.6 Facts from Linear Algebra

- $\lambda \in \mathbb{C}$ is an *eigen-value* and the non-zero vector $v \in \mathbb{C}^n$ is an *eigen-vector* of the $n \times n$ matrix A if $Av = \lambda v$. If v is an eigen-vector of A then so is any non-zero scalar multiple of v.

- Every square matrix has a *Jordan canonical form* $A = PJP^{-1}$ where P is a non-singular 'change of basis' and J is block diagonal with Jordan blocks (see section 2.1).

- If the square matrix A hs a complete set of eigen-vectors then all of its Jordan blocks are 1×1 and it is 'similar' to a diagonal matrix Λ: $A = P\Lambda P^{-1}$. The columns of P are the eigen-vectors of A and the diagonal elements of Λ are the eigen-values of A.

- NOTICE that the equation $A = P\Lambda P^{-1}$ holds if and only if $AP = P\Lambda$, and this equation is precisely $A\mathbf{p}_j = \mathbf{p}_j\lambda_j$ for $j = 1, 2, ..., n$ where $\mathbf{p}_j$ is the j-th column of P and λ_j is the j-th eigen-value:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

  or

$$A \begin{bmatrix} \mathbf{p}_1 \, \mathbf{p}_2 \cdots \mathbf{p}_n \end{bmatrix} = \begin{bmatrix} \mathbf{p}_1 \, \mathbf{p}_2 \cdots \mathbf{p}_n \end{bmatrix} \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \cdots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}$$

- If the $n \times n$ matrix A is real and symmetric, $A' = A$ (or complex and Hermitian, $A^* = A$), then A will be diagonizable (be similar to a diagonal matrix Λ). In this case it will have all real eigen-values and a complete set of orthonormal eigen-vectors in $\mathbb{R}^n$ (or $\mathbb{C}^n$). Thus,

$$A = P\Lambda P' \qquad (\text{or} \quad A = P\Lambda P^*)$$

  where $P' = P^{-1}$ (or $P^* = P^{-1}$).

- If A and B are $n \times n$, $\det(AB) = \det(A)\det(B)$. So $\det(A)\det(A^{-1}) = \det(I) = 1$, and therefore if $A = P\Lambda P^{-1}$ we have $\det(A) = \det(\Lambda) = \lambda_1 \cdots \lambda_n$.

- A symmetric (Hermitian) matrix A is *positive definite* if $x'Ax > 0$ for all non-zero $x \in \mathbb{R}^n$ ($z^*Az > 0$ for all non-zero $z \in \mathbb{C}^n$). And this condition holds if and only if all eigen-values of A are positive.

  A symmetric (Hermitian) matrix A is *positive semi-definite* if $x'Ax \geq 0$ for all $x \in \mathbb{R}^n$ ($z^*Az \geq 0$ for all $z \in \mathbb{C}^n$). And this condition holds if and only if all eigen-values of A are ≥ 0.

- Every $m \times n$ matrix A has $\min\{m, n\}$ *singular values*. If $m \leq n$, these are the eigen-values, $\sigma_1, \sigma_2, ..., \sigma_m$, all necessarily ≥ 0, of the positive semi-definite matrix AA^*; or if $n \leq m$ of A^*A.

  Every $m \times n$ matrix A, then, has a *singular value decomposition* (SVD): $A = U\Sigma V$, where Σ is an $m \times n$ matrix of zeros except for the singular values, the σ_j's, down the diagonal, the columns of the $m \times m$ matrix U are the eigen-vectors of AA^*, and the columns of the $n \times n$ matrix V are the eigen-vectors of A^*A.

Positive definite matrices are important in applications because (i) they give us the metrics for various coordinate systems in $\mathbb{R}^n$ (or $\mathbb{C}^n$), (ii) the Gram matrices that arise in Hilbert space projections are positive definite, and (iii) covariance matrices for vectors of random variables are positive definite.

1 Function Spaces and Linear Transformations

1.1 Vector Spaces, Norms, Inner Products, Limits

Vector spaces $\mathbb{K}$ will always denote the real $\mathbb{R}$ or complex $\mathbb{C}$ numbers.

1.1.1 Definition. A *vector space* or *linear space* over the field $\mathbb{K}$ is a set X which is an abelian group under an operation called vector addition, denoted $+$, and which has a scalar multiplication that is distributive over vector addition. To be specific, for any x, y, and z in X, and any α and β in $\mathbb{K}$ the vector addition $x+y$ and the scalar multiplication αx satisfy:

a) $x+y \in X$;

b) $(x+y)+z = x+(y+z)$;

c) $x+y = y+x$;

d) there exists $0 \in X$ such that $0+x = x$;

e) there exists $-x \in X$ such that $x+(-x) = 0$, the left we denote by $x-x$;

f) $\alpha x \in X$;

g) $1\,x = x$;

h) $0\,x = 0$; (on the left $0 \in \mathbb{K}$, on the rght $0 \in X$)

i) $(\alpha\beta)x = \alpha(\beta x)$;

j) $\alpha(x+y) = \alpha x + \alpha y$;

k) $(\alpha+\beta)x = \alpha x + \beta x$.

Here are some examples of vector spaces. The reader should mentally check some points of the definition for each case.

- For any $n \in \mathbb{N}$, the familiar $\mathbb{K}^n$ ($\mathbb{R}^n$ or $\mathbb{C}^n$) is a vector space.
- The set of infinite (real or complex) sequences, which we arrange in a 'tuple' $(x_1, x_2, x_3, ...)$, is a vector space with addition and scalar multiplication given by

$$(x_1, x_2, x_3, ...) + (y_1, y_2, y_3, ...) = (x_1+y_1, x_2+y_2, x_3+y_3, ...) \qquad \text{and}$$
$$\alpha(x_1, x_2, x_3, ...) = (\alpha x_1, \alpha x_2, \alpha x_3, ...)\,.$$

- For any $n \in \mathbb{N}$, the set of polynomials in one variable of degree less than or equal to n is a vector space. So is the set of (at most) n-th degree polynomials in d variables. These have the form

$$p(x_1, ..., x_d) \;=\; \sum_{\alpha_1+\cdots+\alpha_d \le n} a_{(\alpha_1,...,\alpha_d)} x_1^{\alpha_1} \cdots x_d^{\alpha_d} \qquad \text{or} \qquad p(x) = \sum_{|\alpha| \le n} a_\alpha x^\alpha$$

  for short.

- The set of all polynomials in one variable, of any degree, is a vector space. And so is the set of polynomials, of any degree, in d variables.
- The set $C((0,1))$ of continuous, scalar valued, functions on the open interval $(0,1)$, is a vector space.
- The set $C([0,1])$ of continuous, scalar valued, functions on the closed interval $[0,1]$, is a vector space. Every function in this vector space has one sided limits at the end points.

  The set $C([0,1])$ is a subset of $C((0,1))$; the later includes functions like $f(x) = 1/(1-x)$ but the former does not.

- The set of square integrable functions on (0.1), $L^2(0,1)$, is a vector space. These functions satisfy

$$\int_0^1 |f(x)|^2, dx \ < \ \infty\,.$$

  The L denotes the Lebesgue integral, and a precise definition of this set will be given later.

  It is not immediately clear that this set of functions is closed under addition of functions. This will also be shown later (in section 1.3).

If V is a subset of a vector space X, and if V is itself is a vector space, we say V is a *sub-vector space* or *subspace* of X. For fixed $n \in \mathbb{N}$, the polynomials of degree at most n are a subspace of the vector space of all polynomials of any degree. And the set of all polynomials on $\mathbb{R}$, when restricted to $[0,1]$, is a subspace of $C([0,1])$, which is itself a subspace of $C((0,1))$.

If U and V are both subspaces of a vector space X then the intersection of these two sets $U \cap V$ is also a sub-vector space (exercise 1.1.21). The intersection is never empty for it necessarily contains $0 \in X$.

In general the union of two subspaces of a vector space is not a subspace. However, the 'sum' of two subspaces, $U + V = \{x = u + v \ ; \ u \in U \text{ and } v \in V\}$, is a subspace.

1.1.2 Lemma. *If U and V are subspaces of the vector space X and if $U \cap V = \{0\}$ then every $x \in U + V$ has a unique representation as $x = u + v$ with $u \in U$ and $v \in V$.*

Proof. Since $x \in U + V$ is it obvious that $x = u + v$ for *some* $u \in U$ and $v \in V$. We wish to show that u and v are unique. Suppose also that $x = u' + v'$. Then $0 = x - x = (u + v) - (u' + v') = (u - u') + (v - v')$ where $(u - u') \in U$ (since U is a subspace) and $(v - v') \in V$. But this equation also says (since 0 belongs to both U and V) that $u - u' = 0 - (v - v')$ which is a vector in V since the right hand side is. Thus $u - u' \in U \cap V$ or $u - u' = 0$. Similarly, $v - v' = 0 - (u - u') \in U$ as well as V, and therefore $v - v' = 0$. □

A finite collection of vectors $x_1, x_2, ..., x_n$ in a vector space X is *linearly independent* if the condition $\alpha_1 x_1 + \cdots + \alpha_n x_n = 0$ in X implies $\alpha_1 = \cdots = \alpha_n = 0$ in $\mathbb{K}$ whenever $\alpha_1, ..., \alpha_n$ are scalars in $\mathbb{K}$. Intuitively, a set of vectors is linearly independent if each vector in the set points in a different direction. A set of vectors which is not linearly independent is *linearly dependent*. A finite collection of vectors $x_1, x_2, ..., x_n$ in a vector space X is a *spanning set* for X, or *spans* X, if every vector $x \in X$ may be written $x = \alpha_1 x_1 + \cdots + \alpha_n x_n$ for some scalars $\alpha_1, ..., \alpha_n$. A set of vectors $x_1, x_2, ..., x_n$ in X which is linearly independent and which spans X is a *basis* for X. If two finite sets $\{x_1, ..., x_m\}$ and $\{y_1, ..., y_n\}$ are both bases for the same vector space X, it can be shown that $m = n$, i.e., the cardinality of both sets is the same. This number, the number of vectors in any basis for X, is the dimension of X. If there is no finite set which spans X, X is said to be *infinite dimensional*.

The set of monomials $\{1, x, x^2, ..., x^n\}$ is a basis for the $n+1$-dimensional vector space of polynomials of degree $\leq n$. The vector spaces of all polynomials on $\mathbb{R}$, $C([0,1])$, and $C((0,1))$ are infinite dimensional.

1.1.3 Exercise. Let $[a,b]$ be a closed bounded interval in $\mathbb{R}$ and denote by $\mathscr{R}([a,b])$ the rational function on $[a,b]$, that is, the set of $f(x) = p(x)/q(x)$ such that p and q are polynomials of x and q has no roots in $[a,b]$. Prove that $\mathscr{R}([a,b])$ is a vector space.

Let $K \subset \mathbb{R}^d$ be compact. Define $\mathscr{R}(K)$ and prove it is a vector space.

1.1.4 Exercise. Let $[a,b]$ be a closed bounded interval in $\mathbb{R}$ and denote by $C([a,b];\mathbb{R}^n)$ the set of continuous functions on $[a,b]$ which are valued in $\mathbb{R}^n$. Prove that $C([a,b];\mathbb{R}^n)$ is a vector space.

Let $K \subset \mathbb{R}^d$ be compact. Define $C(K;\mathbb{R}^n)$ and prove it is a vector space.

1.1.5 Exercise. Let $[a,b]$ be a closed bounded interval in $\mathbb{R}$ and denote by $C^1([a,b])$ the set of continuous functions on $[a,b]$ which are continuously differentiable on $[a,b]$. (Assume the necessary one-sided limits exist at a and b.) Prove that $C^1([a,b])$ is a vector space.

Define $C^1([a,b];\mathbb{R}^n)$ and prove it is a vector space.

1.1.6 Exercise. Prove that both $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ are vector spaces, over $\mathbb{R}$ and $\mathbb{C}$ respectively.

Metric spaces

1.1.7 Definition. A *metric space* is a set X together with a function $d : X \times X \to \mathbb{R}$ which satisfies the following properties:

a) $d(x, y) \geq 0$ for all x and $y \in X$;

b) $d(x, y) = 0$ if and only if $x = y$ in X;

c) $d(x, y) = d(y, x)$ for all x and $y \in X$;

d) $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in X$ (triangle inequality).

The function d is called a *metric* or *distance function.*

In analysis, a metric is vital for defining open and closed sets, convergence, continuity, and many other concepts. This section gives a short review.

Here are some examples of metrics on the vector spaces that we listed in the last paragraph.

- The familiar vector length $|x| = \sqrt{x_1^2 + \cdots + x_n^2}$ gives rise to the metric $d(x, y) = |x - y|$ on $\mathbb{R}^n$, for any $n \in \mathbb{N}$. On $\mathbb{C}^n$ the vector length that must be used here is $|z| = \sqrt{|z_1|^2 + \cdots + |z_n|^2}$. When $n = 1$, $|z|^2 = |x + iy|^2 = x^2 + y^2$.

- Consider the set of infinite (real or complex) tuples $x = (x_1, x_2, x_3, ...)$ which satisfy

$$\|x\|^2 \;=\; \sum_{j=1}^{\infty} |x_j|^2 \;\leq \infty\,.$$

  This set is a vector space which we denote by $\ell^2(\mathbb{N})$.

  It is not obvious that $\ell^2(\mathbb{N})$ is closed under the vector addition that we already defined for infinite tuples. This fact follows from two inequalities: $2ab \leq a^2 + b^2$ for every $a, b \in \mathbb{R}$ (which holds because $0 \leq (a-b)^2$), and $|\sum_{j=1}^{\infty} x_j \bar{y}_j|^2 \leq \sum_{j=1}^{\infty} |x_j|^2 \sum_{j=1}^{\infty} |y_j|^2$ which is Schwarz inequality (already familiar in $\mathbb{R}^n$) and will be proven later. Using these we have

$$\begin{aligned}\sum_{j=1}^{\infty} |x_j + y_j|^2 \;&=\; \sum_{j=1}^{\infty} \big[|x_j|^2 + 2\Re(x_j\bar{y}_j) + |y_j|^2\big] \;\leq\; \sum_{j=1}^{\infty} \big[|x_j|^2 + 2|x_j\bar{y}_j| + |y_j|^2\big] \\ &\leq\; \sum_{j=1}^{\infty} \big[|x_j|^2 + (|x_j|^2 + |y_j|^2) + |y_j|^2\big] \;=\; 2\sum_{j=1}^{\infty} \big[|x_j|^2 + |y_j|^2\big]\,.\end{aligned}$$

  Thus, if $(x_1, x_2, x_3, ...)$ and $(y_1, y_2, y_3, ...)$ belong to $\ell^2(\mathbb{N})$, so does $(x_1 + y_1, x_2 + y_2, x_3 + y_3, ...)$.

  A metric on $\ell^2(\mathbb{N})$ is now defined by $d(x, y) = \|x - y\| = \sqrt{\sum_{j=1}^{\infty} |x_j - y_j|^2}$.

- The vector 'length' function $\|f\| = \sup\{|f(x)| \,;\, 0 \leq x \leq 1\}$ gives rise to the metric

$$d(f, g) \;=\; \|f - g\| \;=\; \sup\{|f(x) - g(x)| \,;\, 0 \leq x \leq 1\}$$

  on $C([0, 1])$.

  It also is a metric on the vector space of all polynomials on $\mathbb{R}$.

- Another vector 'length' that may be used on $C([0, 1])$ is

$$\|f\| \;=\; \sqrt{\int_0^1 |f(x)|^2\, dx}\,.$$

  Again, the corresponding metric is $d(f, g) = \|f - g\| = \sqrt{\int_0^1 |f(x) - g(x)|^2\, dx}$.

- The vector space $C((0,1))$ may also be given a metric but it is rather technical to write down and we will not do so here. (There is no vector length or 'norm' which can be used to define a metric here.)

- A subset of a metric space is itself a metric space using the same metric. Any subset (not necessarily a subspace) of the vector space $\mathbb{R}^n$ or $C([0,1])$, with metric given above, is a metric space. For instance, if $r > 0$ is fixed the set $\{f(x) \in C([0,1]) \; ; \; -r < |f(x) \leq r \text{ for all } x \in [0,1]\}$ is a metric space.

On any metric space we define the (open) ball, centered at $x \in X$ and of radius $r > 0$ by

$$B(x,r) \; = \; \{y \in X \; ; \; d(x,y) < r\} \, .$$

Here are some other definitions and facts that will be routinely used.

Let X be a metric space with distance function $d(\cdot)$. A set $A \subset X$ is *open* if it contains an open ball about each of its points; that is, A is open if for every $x \in A$ there is an $r > 0$ such that $B(x,r) \subset A$. It is easy to see (just check the definition of open) that an arbitray union (finite, countably infinite, or uncountable) of open sets is open. For any subset $A \subset X$ we may consider the collection of all open sets contained in A. The union of this collection is both open and contained in A. It is the largest open set contained in A and is called the *interior* of A, denoted A^o or $\text{int}(A)$.

A set $A \subset X$ is *closed* if its complement, $A^c = X \backslash A$, is open. By taking complements of arbitrary unions, it is easy to see that an arbitray intersection of closed sets is closed. For any subset $A \subset X$ we may consider the collection of all closed sets that contain A. The intersection of this collection is both closed and contains A. It is the smallest closed set that contains A and is called the *closure* of A, denoted $\overline{A}$ or $\text{cl}(A)$.

The *boundary* of a subset $A \subset X$, denoted ∂A or $\text{bdy}(A)$, is the intersection of its closure and the closure of its complement. That is, $\partial A = \overline{A} \cap \overline{A^c}$.

A set $A \subset X$ is *bounded* if there is a point $x \in X$ and an $R > 0$ so large that $A \subset B(x,R)$. Note that if A has this porperty for some $x \in X$, then it has this property, perhaps with a larger R, for any other $x' \in X$ (since $d(x,x')$ is finite, and using the triangle inequality).

Compactness is an important property in analysis. A general, and useful, definition is that a set is compact if every open cover is reducible to a finite sub-cover. However, in metric spaces the following definition is equivalent: $A \subset X$ is *compact* if every sequence of points x_n in A has a sub-sequence x_{n_k} which converges to a point $x \in A$ (as $k \to \infty$). (This is the Bolzano-Weierstrass Theorem, and this definition is sometimes called sequential compactness.) When X is a subset of $\mathbb{R}^n$ or $\mathbb{C}^n$, it can be shown that $A \subset X$ is compact if and only if A is closed and bounded. (This is the Heine-Borel Theorem.)

A sequence x_n in X *converges* to a point, a *limit*, x in X if for every $\epsilon > 0$ there is an $n_0 \in \mathbb{N}$ such that $d(x,x_n) < \epsilon$ whenever $n \geq n_0$.

The concept of Cauchy sequence allows us to determine if a sequence 'wants' to converge without the need to make explicit reference to its limit. A sequence x_n in X is a *Cauchy sequence* if for every $\epsilon > 0$ there is an $n_0 \in \mathbb{N}$ such that $d(x_m,x_n) < \epsilon$ whenever both $m \geq n_0$ and $n \geq n_0$. Every convergent sequence is Cauchy (the reader should be able to prove this using the triangle inequality: $d(x_m,x_n) \leq d(x_m,x) + d(x,x_n)$).

A metric space X is *complete* if every Cauchy sequence x_n in X has a limit $x = \lim_{n\to\infty} x_n$ in X. The reals are complete, the rationals are not.

Thus, in a complete metric space, as sequence converges if and only if it is Cauchy. The reader may also remember, and perhaps be able to prove, that a closed subset of a complete metric space is itself complete.

A subset $A \subset X$ is *dense* in X if it has the following property: for every $x \in X$ and every $\epsilon > 0$ there is a $y \in A$ such that $d(x,y) < \epsilon$. The rationals are dense in the reals. And the polynomials are dense in the space of continuous functions on $[0,1]$ when the 'sup-norm' (uniform convergence) is used; this is the Weierstrass approximation theorem.

A metric space X is *separable* if it has a countable dense subset. The reals are separable because the rationals are both dense in $\mathbb{R}$, and countable. The reader should be able to use this fact to show that both $\mathbb{R}^n$ and $\mathbb{C}^n$ are separable.

If X and Y are metric spaces, a function $f : X \to Y$ is *continuous* if it has any one of the following equivalent properties:

- $f^{-1}(B)$ is open in X whenever B is an open subset of Y.

- For every $x \in X$ and $\epsilon > 0$ there is a $\delta > 0$ such that $x, x' \in X$ and $d(x, x') < \delta$ implies $d(f(x), f(x')) < \epsilon$ in Y.
- For every sequence x_n in X with limit $x = \lim_{n\to\infty} x_n$ in X, the sequence $f(x_n)$ in Y has limit $f(x) = \lim_{n\to\infty} f(x_n)$ in X. That is, f is continuous if the interchange of limits $\lim f(x_n) = f(\lim x_n)$ always holds.

We have not done so, but the reader should be able to define continuity at a single point $x \in X$.

The following theorem is constantly used in analysis. It implies that, whenever necessary, we may as well assume that our metric spaces are complete.

1.1.8 Theorem. *Every metric space has a unique completion. More precisely, if X is any metric space, there is a complete metric space $\overline{X}$ in which X is dense.*

Solutions of complex problems in modern analysis are often given by constructing an approximating sequence, then showing that this sequence is Cauchy in a complete metric space.

Norms

1.1.9 Definition. A *normed vector space*, or just *normed space*, is a vector space X together with a function $\|\cdot\| : X \to \mathbb{R}$ which satisfies the following properties:

a) $\|x\| \geq 0$ for all $x \in X$;

b) $\|x\| = 0$ only when $x = 0$;

c) $\|\alpha x\| = |\alpha|\,\|x\|$ for all $\alpha \in \mathbb{K}$ and $x \in X$;

d) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$.

The function $\|\cdot\|$ is called a *norm*.

Since

$$\|x - z\| \;=\; \|x - y + y - z\| \;\leq\; \|x - z\| + \|z - y\| \qquad \text{for every} \qquad x, y, z \in X\,,$$

the triangle inequality holds for the function $d(x, y) = \|x - y\|$. It is then easy to see that *every normed vector space is a metric space* with this distance function.

Examples of normed vector spaces are given by the first four bullets in the list of examples of metric spaces. These are $\mathbb{R}^n$, $\mathbb{C}^n$, $\ell^2(\mathbb{N})$ (either real of complex valued sequences), and $C([0, 1])$ (either real or complex valued functions) with two possible norms.

Different norms may be used to define the same open sets, Cauchy and convergent sequences, continuous functions, etc. The vector spaces $\mathbb{K}^n$ are normed vector spaces with any of the norms:

- $\|x\| = (\sum_{i=1}^n |x_i|^2)^{1/2}$,
- $\|x\| = \sum_{i=1}^n |x_i|$,
- $\|x\| = \sup_{1\leq i\leq n} |x_i|$.

Here, $x = (x_1, ..., x_n)$ and $|x_i|$ is the absolute value or modulus of the real or complex number x_i. Throughout this manuscript the first of the above norms, the familiar Euclidean length, will be denoted by $|x|$, the same notation used for the scalar absolute value in the special case $n = 1$. Thus, for $x = (x_1, ...x_n) \in \mathbb{R}^n$, $|x| = (\sum_{j=1}^n x_j^2)^{1/2}$, and when $z = (z_1, ...z_n) = (x_1, ...x_n) + i(y_1, ...y_n) \in \mathbb{C}^n$,

$$|z| \;=\; \Big(\sum_{j=1}^n \overline{z}_j z_j\Big)^{1/2} \;=\; \Big(\sum_{j=1}^n |z_j|^2\Big)^{1/2} \;=\; \Big(\sum_{j=1}^n \; x_j^2 + y_j^2 \;\Big)^{1/2}\,.$$

A norm can be put on the (infinite dimensional) vector space of polynomials (in one variable) by setting

$$\|f\| \;=\; \sup\{|f(x)| \;;\; 0 \leq x \leq 1\} \tag{1.1.1}$$

when f is any polynomial. This number, $\|f\|$, is certainly finite for every polynomial f. It is easy to see that $\|f+g\| \le \|f\| + \|g\|$ for any two polynomials f and g.

Let $C([0,1])$ denote the vector space of functions which are continuous on the compact set $[0,1]$. Such functions always assume their maximum and minimum on $[0,1]$ (since the continuous image of a compact set is compact). So $\|f\| < \infty$ for every $f \in C([0,1])$.

In this space $f_n \to f$ means that $\|f_n - f\| \to 0$ as $n \to \infty$. If $f_n \to f$ in this norm, we say that 'f_n converges *uniformly* to f as $n \to \infty$'. We leave it for the reader to show (exercise 1.1.29) that (1.1.1) is indeed a norm.

That $C([0,1])$ is a vector space follows from elementary facts about continuous functions: The sum of two continuous functions is continuous as are scalar multiples of such functions. The other vector space properties are also easy to check.

It is shown in elementary real variables courses that a uniformly convergent sequence of continuous functions always converges to a continuous function. (This is done with an "$\epsilon/3 + \epsilon/3 + \epsilon/3$ argument".) This argument proves the following

1.1.10 Theorem. *The vector space $C([0,1])$ is complete when given the norm $\|\cdot\|$ defined by (1.1.1).*

The norm $\|\cdot\|$ on any normed vector space X is a continuous function when X is given the metric $d(\cdot,\cdot)$ induced by this norm. It suffices to show that for every $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\|x - y\| < \delta \qquad \text{implies} \qquad |\,\|x\| - \|y\|\,| < \epsilon.$$

To this end observe that $\|x\| \le \|x - y\| + \|y\|$ so $\|x\| - \|y\| \le \|x - y\|$. Similarly (interchange the roles of x and y) $\|y\| - \|x\| \le \|y - x\| = \|x - y\|$. Putting these two expressions together we conclude that

$$|\|x\| - \|y\|| \le \|x - y\|.$$

This shows that given ϵ we may take $\delta = \epsilon$. In fact we have shown that $\|\cdot\|$ is uniformly continuous, and Lipschitz continuous (with Lipschitz constant 1), on X.

Translating the definitions of convergent and Cauchy sequences into the present context of a normed space, a sequence x_1, x_2, x_3, ... in a normed vector space X *converges* to $x \in X$ if $\|x_n - x\| \to 0$ as $n \to \infty$. We again use the notation $x_n \to x$ as $n \to \infty$, or $\lim_{n\to\infty} x_n = x$ for a convergent sequence. A sequence x_1, x_2, x_3, ... in X is *Cauchy* if for every $\epsilon > 0$ there exists an $n_0 \in \mathbb{N}$ such that $n, m \ge n_0$ implies $\|x_m - x_n\| < \epsilon$. And X is *complete* if every Cauchy sequence converges in X.

1.1.11 Definition. A complete normed vector space is called a *Banach space*.

The normed vector spaces $\mathbb{K}^n$, $\ell^2(\mathbb{N})$, and $C([0,1])$ are Banach spaces.

The vector space of polynomials restricted to $[0,1]$, with the $C([0,1])$ norm (1.1.1), is not complete. In fact the Weierstrass approximation theorem states that these polynomials are dense in $C([0,1])$. So $C([0,1])$ is the completion of the polynomials on $[0,1]$.

Sequences can be defined in any metric space; the additional structure of a vector space allows us to define *infinite series* as well. In a normed vector space the notions of sequence and series are the same in the following sense. The partial sums of a series, $s_n = \sum_{j=1}^{n} x_j$, are the terms of a sequence. And every sequence s_n can be expressed as the partial sums of the series of its differences $s_n = \sum_{j=1}^{n} x_j$ where $x_1 = s_1$ and $x_j = s_j - s_{j-1}$ when $j \ge 2$.

The following proposition gives a useful sufficient, but not necessary, condition for convergence of a series in a Banach space.

1.1.12 Proposition (Weierstrass M-test)**.** *Let X be a Banach space and x_n a sequence of elements of X. The infinite series $\sum_1^\infty x_n$ is convergent in X if the series of non-negative reals $\sum_1^\infty \|x_n\|$ is convergent is $\mathbb{R}$. (If $\sum_1^\infty \|x_n\| < \infty$ we say the series converges absolutely.)*

Proof. The limit $\lim_{n\to\infty} \sum_1^n x_k$ exists if and only if the sequence of partial sums is Cauchy in X. For any $m \ge n$ in $\mathbb{N}$ we have the bound

$$\|\sum_{k=n}^{m} x_k\| \le \sum_{n}^{m} \|x_k\|\,.$$

This shows that $\sum_1^n x_k$ is Cauchy in X whenever $\sum_1^n \|x_k\|$ is Cauchy in $\mathbb{R}$. □

Inner products We now add the useful notion of orthogonality between vectors.

1.1.13 Definition. An *inner product space* is a vector space X over a field $\mathbb{K}$ ($= \mathbb{C}$ or $\mathbb{R}$) together with a function $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{K}$ such that

(a) $\langle x, x \rangle \geq 0$ for all $x \in X$.

(b) $\langle x, x \rangle = 0$ implies $x = 0$ in X.

(c) $\langle y, x \rangle = \overline{\langle x, y \rangle}$ when $\mathbb{K} = \mathbb{C}$ and $x, y \in X$, (or $\langle y, x \rangle = \langle x, y \rangle$ when $\mathbb{K} = \mathbb{R}$).

(d) $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$ for all $x, y, z \in X$.

(e) $\langle x, \alpha y \rangle = \alpha \langle x, y \rangle$ for all $x, y \in X$ and $\alpha \in \mathbb{K}$.

The function $\langle \cdot, \cdot \rangle$ is called an *inner product* or *scalar product.*

Here are some additional properites of the inner product. Using $\alpha = 0$ in (e) shows that

(f) $\langle x, 0 \rangle = 0$ for all $x \in X$.

Using (c) with (e) shows that

(g) $\langle \alpha x, y \rangle = \overline{\alpha} \langle x, y \rangle$ when $\mathbb{K} = \mathbb{C}$, or $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$ when $\mathbb{K} = \mathbb{R}$.

Combining (c) and (d) shows that

(h) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$.

Here are some examples of inner product spaces.

- $\mathbb{R}^n$ and $\mathbb{C}^n$ with $\langle x, y \rangle = \sum_{j=1}^n \bar{x}_j y_j$. (Thinking of x and y as column vectors, we will usually write this inner product as $x^* y$, or $x' y$ in the real case.)
- $\ell^2(\mathbb{N})$ with inner product $\langle x, y \rangle = \sum_{j=1}^\infty \bar{x}_j y_j$. (Omit conjugate in the real case.)
- $C([0,1])$ with inner product $\langle f, g \rangle = \int_0^1 \bar{f}(x) g(x)\, dx$. (Omit conjugate in the real case.)

1.1.14 Definition. Let $x_1, x_2, ..., x_n$ be vectors in a real or complex vector space with inner product $\langle \cdot, \cdot \rangle$. Then the $n \times n$ matrix

$$G \;=\; \begin{pmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \cdots & \langle x_1, x_n \rangle \\ \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & \cdots & \langle x_2, x_n \rangle \\ \vdots & \vdots & & \vdots \\ \langle x_n, x_1 \rangle & \langle x_n, x_2 \rangle & \cdots & \langle x_n, x_n \rangle \end{pmatrix} \tag{1.1.2}$$

is called the *Gram matrix* for $x_1, x_2, ..., x_n$.

1.1.15 Proposition. *The Gram matrix G is symmetric, $G' = G$, when $\mathbb{K} = \mathbb{R}$; G is Hermitian, $G^* = G$, when $\mathbb{K} = \mathbb{C}$. G is always positive semi-definite.*[2] *And G is strictly positive definite if and only if the vectors $x_1, x_2, ..., x_n$ are linearly independent.*

Proof. By the definition of inner product $\langle x_i, x_j \rangle = \overline{\langle x_j, x_i \rangle}$. Since each side is the ij-th entry of G and G^*, respectively, the first statement is true.

Next let $\alpha_1, ..., \alpha_n$ be scalars, let a denote the column vector whose j-th row is $\overline{\alpha}_j$, and set $y = \alpha_1 x_1 + \cdots + \alpha_n x_n$. The properties of the inner product show that

$$0 \;\leq\; \langle y, y \rangle \;=\; \sum_{i=1}^n \sum_{j=1}^n \alpha_i \overline{\alpha}_j \langle x_i, x_j \rangle \;=\; a^* G a \;. \tag{1.1.3}$$

Therefore G is positive semi-definite.

Finally, let y and a be as before. First assume the x_j's are linearly independent and $a \neq 0$ in $\mathbb{K}^n$. Then $y \neq 0$ and the first inequality in (1.1.3) is strict. So G is strictly positive definite. Conversely, assume that G is strictly positive definite. Setting $y = 0$ in (1.1.3) then implies that $a^* M a = 0$, and hence $a = 0$. Thus the x_j's are linearly independent. □

[2] A matrix G is positive semi-definite if $x' G x \geq 0$ (or $x^* G x \geq 0$ in the complex case) for every vector x. It is (strictly) positive definite if $x' G x > 0$ (or $x^* G x > 0$ in the complex case) for every vector $x \neq 0$. G is positive (semi-) definite if and only if every principle sub-matrix of G has determinant > 0 (≥ 0). A principle sub-matrix is one which is itself on the diagonal of G, i.e., of the form $(a_{ij})_{k \leq i,j \leq \ell}$ for some $k < \ell$. And G is positive (semi-) definite if and only if every eigenvalue of G is (real and) positive (non-negative).

1.1.16 Proposition (Schwarz inequality)**.** *If X is an inner product space and we define the function $\|x\| = \langle x, x\rangle^{1/2}$ for $x \in X$, then*

$$|\langle x, y\rangle| \le \|x\|\,\|y\| \tag{1.1.4}$$

for all $x, y \in X$. The inequality is strict if and only if x and y are linearly independent.

Proof. Pick any $x, y \in X$. If $\langle x, y\rangle = 0$ the inequality to be shown is obvious; so consider the case $\langle x, y\rangle \neq 0$. Set $\alpha = |\langle x, y\rangle|/\overline{\langle x, y\rangle} \in \mathbb{C}$. Clearly $|\alpha| = 1$. So for any $r \in \mathbb{R}$ we have

$$\begin{aligned} 0 \le \|r\alpha x - y\|^2 &= \langle r\alpha x - y, r\alpha x - y\rangle = r^2|\alpha|^2\|x\|^2 - r\overline{\alpha}\langle x, y\rangle - r\alpha\overline{\langle x, y\rangle} + \|y\|^2 \\ &= r^2\|x\|^2 - 2r|\langle x, y\rangle| + \|y\|^2 = ar^2 - 2br + c \end{aligned}$$

where $a = \|x\|^2$, $b = |\langle x, y\rangle|$, and $c = \|y\|^2$ are all real.

Now the minimum of this quadratic occurs when the derivative is zero: $2ar - 2b = 0$ or $r = b/a$. Substituting back into the quadratic gives

$$0 \le a(b^2/a^2) - 2b(b/a) + c = -b^2/a + c\,, \qquad \text{or} \qquad b^2/a \le c\,.$$

But this means $b^2 \le ac$ which is the inequality to be shown.

Note that the initial inequality $0 \le \|r\alpha x - y\|^2$ may be replaced by $0 = \|r\alpha x - y\|^2$, and so in all subsequent occurrences, if and only if $r\alpha x = y$. □

1.1.17 Proposition. *If X is an inner product space the function $x \mapsto \|x\| \stackrel{def}{=} \langle x, x\rangle^{1/2}$ is a norm which makes X into a normed vector space.*

Proof. That the range of $\|\cdot\|$ is contained in $[0, \infty)$ follows from (a) in Definition 1.1.13. Property (b) in Definition 1.1.9 is immediate from Definition 1.1.13 (b). Using Definition 1.1.13 (c) and (e) gives

$$\|\alpha x\|^2 = \langle \alpha x, \alpha x\rangle = \overline{\alpha}\langle x, \alpha x\rangle = \overline{\alpha}\alpha\langle x, x\rangle = |\alpha|^2\|x\|^2$$

which shows Definition 1.1.9 (c). To show Definition 1.1.9 (d) expand

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y\rangle = \langle x, x\rangle + \langle x, y\rangle + \langle y, x\rangle + \langle y, y\rangle \\ &= \langle x, x\rangle + \langle x, y\rangle + \overline{\langle x, y\rangle} + \langle y, y\rangle = \|x\|^2 + 2\,\Re\langle x, y\rangle + \|y\|^2 \\ &\le \|x\|^2 + 2|\langle x, y\rangle| + \|y\|^2 \le \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2 \\ &= (\|x\| + \|y\|)^2. \end{aligned}$$

The first inequality holds because the real part of the complex number $\langle x, y\rangle$ is bounded by its modulus, that is, $u \le \sqrt{u^2 + v^2}$ for any $w = u + iv \in \mathbb{C}$; the second is Schwarz' inequality. □

Proposition 1.1.17 implies that *every inner product space is a normed vector space* with norm $\|x\| = \langle x, x\rangle^{1/2}$.

1.1.18 Definition. An inner product space that is complete with respect to the norm induced by its inner product is called a *Hilbert space*.

Every Hilbert space is also a Banach space using the norm defined from its inner product.

The spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ are Hilbert spaces, and so is $\ell^2(\mathbb{N})$ with either real or complex valued sequences.

The set $C([0, 1])$ with inner product defined by $\langle f, g\rangle = \int_0^1 \overline{f}g\,dx$ is not complete. For instance, if $n \in \mathbb{N}$ set $f_n(x) = (x + 1/2)^n$ for $0 \le x \le 1/2$ and $= 1$ for $1/2 < x \le 1$. As $n \to \infty$, $\int_0^1 |f_n(x) - f(x)|^2\,dx \to 0$ where $f(x) = 0$ for $0 \le x < 1/2$ and $= 1$ for $1/2 \le x \le 1$. The sequence f_n is in $C([0, 1])$, and it converges in the norm induced by the inner product (to $f(x)$); but its limit is not in $C([0, 1])$.

Warning. We will define and make extensive use of the vector space of square integrable functions, $L^2(\Omega)$, and certain subspaces. We use the norm $f \mapsto \sqrt{\int_\Omega |f(x)|^2\,dx}$, which is not capable of distinguishing functions on Ω which are equal almost everywhere. The statement $f = g$ in $L^2(\Omega)$ no longer means $f(x) = g(x)$ for all $x \in \Omega$; it means $\int_\Omega |f(x) - g(x)|^2\,dx = 0$, and the expression $f = g$ is preferable to $f(x) = g(x)$.

Warning. It is common to use several vector spaces in the course of analyzing a single problem, and it can be tempting to use symbols like + and = carelessly. When rigor is at stake one should be alert to ask what is the common vector space or set containing x and y that allows us to write $x + y$ or $x = y$.

We note that a finite dimensional subspaces of any normed vector space X,whether X is complete or not, is always complete. A finite dimensional subspace is of course isomorphic as a vector space to $\mathbb{K}^n$; the point in question is whether a norm on the infinite dimensional vector space could induce a norm on a finite dimensional subspace which is topologically different from the usual Euclidian norm on $\mathbb{K}^n$.

1.1.19 Proposition. *Let M be a finite-dimensional subspace of a normed vector space X over the field $\mathbb{K}$. Then M is closed and complete.*

Proof. A proof of this proposition can be found in Rudin, *Real and Complex Analysis*, section 4.15. We will give the proof when X is an inner product space. This will be the most important case for us.

Let $x_1, x_2, ..., x_n$ be a linearly independent spanning set for M, and $G = [\langle x_i, x_j \rangle]_{n\times n}$ the strictly positive definite Gram matrix for this set of vectors. The quadratic form $\eta^* G\eta$ satisfies

$$\lambda_1|\eta|^2 \;\le\; \eta^* G\eta \;\le\; \lambda_n|\eta|^2 \qquad\qquad \text{for all } \eta \in \mathbb{K}^n\,.$$

(The complex conjugate is superfluous if $\mathbb{K} = \mathbb{R}$.) Here, $\lambda_1 > 0$ is the smallest eigen-value of G and λ_n is the largest. If $y = \eta_1 x_1 + \cdots + \eta_n x_n$ in M these inequalities can be written

$$\lambda_1|\eta|^2 \;\le\; \|y\|^2 \;\le\; \lambda_n|\eta|^2 \qquad \text{and} \qquad \lambda_n^{-1}\|y\| \;\le\; |\eta|^2 \;\le \lambda_1^{-1}\|y\|,.$$

Therefore a sequence in M is Cauchy if and only if the sequence of coefficients in $\mathbb{K}^n$ is Cauchy. Since $\mathbb{K}^n$ is complete the limit there yields n coefficients for a corresponding limiting vector in M. □

We end this section by stating an important corollary of Theorem 1.1.8 in the context of vector spaces with a norm or inner product.

1.1.20 Theorem. *Every normed vector space is a subspace of a Banach space in which it is dense, and every inner product space is a subspace of a Hilbert space in which it is dense.*

To be more precise, if X is a normed vector space there is a Banach space $\overline{X}$ such that $X \subset \overline{X}$ is dense. This $\overline{X}$ is unique up to isomorphism. And if X is an inner product space there is a Hilbert space $\overline{X}$ such that $X \subset \overline{X}$ is dense; this $\overline{X}$ is unique up to isomorphism.

MATCHED FILTERS We end this section with a digression into an important application of inner products. In engineering and science it is often required to process a set of data $\{x_j \;;\; j = 1, 2, ..., n\}$. The problem is to find 'something' of interest in the data, and that 'something' is to be modeled as another set of values $\{a_j \;;\; j = 1, 2, ..., n\}$ which we must compare with the x's.

If the data is numerical we may write $x = (x_1, x_2, ..., x_n)$ and $a = (a_1, a_2, ..., a_n)$ as vectors in $\mathbb{R}^n$ or $\mathbb{C}^n$. Whether the model a is then a good representation of the data x can sometimes be determined by the value of the inner product $a\cdot x = \sum_{j=1}^n a_j x_j$, a measure of how well correlated the two sets of numbers are.

To understand why the inner product is often a good measure of correlation between a and x let's assume that a and x are real and that their lengths are fixed and equal, $|a| = |x| = c > 0$. This is often easily accomplished in a real system by turning the volume, brightness, etc. up or down. We then observe that $a\cdot x$ is maximized, over some set of candidate models $\{a\}$, if and only if the squared length of the difference $|a - x|^2$ is minimized. This is simply because

$$|a - x|^2 \;=\; (a - x)\cdot(a - x) \;=\; |a|^2 - 2\,a\cdot x + |x|^2 \;=\; c^2 - 2\,a\cdot x + c^2\,.$$

Because the values of $|a|$ and $|x|$ are fixed the left side is smallest when $a\cdot x$ is largest. (The reader should interpret this algebra with a visual mental picture of vectors on a sphere of radius c in $\mathbb{R}^3$.)

If the data x is not numerical some other comparison must be used (perhaps just counting the number of components where a and x differ), but it often has some similarities with an inner product in $\mathbb{R}^n$. Whether numerical or not, the comparison of components of two 'vectors,' model and data, is highly parallelizable, and the matched filter is one of the most useful algorithms for extracting information from large data sets.

Here are some examples:

(a) The data is a time series of amplitudes (intensity or energy) of an acoustic or electromagnetic signal and we are looking for the presence of energy at a certain frequency or contained in a certain 'code.'

(b) The data is the greyscale of an image and the model is a picture of a car or some other object of interest. The military uses 'automatic target recognition' algorithms to automatically identify ships, planes, trucks, tanks, or other objects of interest in imagery from satellites and surveillance aircraft. Such satellite 'image processing' is also important to identify foliage, geologic features related to underground water and oil, weather anomalies, and other geographical features of interest.

(c) The data is a text document and the model is a phrase of interest to us, e.g., "to be or not to be." In this setting some conversion into numerical values is sometimes useful, but often a metric is placed on the words or phrases themselves, such as number of letters in which two phrases differ.

(a) A natural model of the signal of interest is

$$a \;=\; (0, \sin\omega\tau, \sin\omega 2\tau, \sin\omega 3\tau, ..., \sin\omega(n-1)\tau)$$

in the case that we are looking for an oscillating function, a signal, at frequency ω (in radians). The parameter τ is the time step size and is determined by the sample rate of an analog to digital converter.

In many applications one cannot be sure that the phase of the data will be aligned with the sinusoid above. A very useful function which accomodates this problem is $e^{i\omega t} = \cos\omega t + i\,\sin\omega t$. We use the inner product in $\mathbb{C}^n$ instead of $\mathbb{R}^n$ with the complex-valued model vector

$$a \;=\; (1, e^{i\omega\tau}, e^{2i\omega\tau}, e^{3i\omega\tau}, ..., e^{(n-1)i\omega\tau})\,.$$

This processing is closely related to the Fourier transform which we will study in chapter 6. The *fast Fourier transform* is a very efficient algorithm for numerical computation of these inner products.

(b) This example naturally is set in $\mathbb{R}^2$ but the $m \times n$ set of pixels is conceptually no different from a one dimensional vector. The inner product between two $m \times n$ real matrices is $A \cdot B = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ij}$.

In both examples (a) and (b) one usually has to search over a large set of data to see what match can be found. For instance searching an image for a 'car' usually means scanning the entire image, and even at many rotated angles, for the set of pixels that look like a car.

(c) There is an entire field of Natural Language Processing (just as there are for signal and image processing) which attempts to automate the process of extracting information from text documents. A nice synopsis can be found on Wikipedia:

`http://en.wikipedia.org/wiki/Natural_language_processing`

Wikipedia has a nice discussion of matched filters:

`http://en.wikipedia.org/wiki/Matched_filter`

1.1.21 Exercise. Let U and V be subspaces of a vector space X. Prove that $U \cap V$ is a vector space. Give an example to show that $U \cup V$ need not be a vector space. Find conditions on U and V which ensure that $U \cup V$ is a vector space. Prove that the set sum $U + V = \{x = u + v\ ;\ u \in U \text{ and } v \in V\}$ is a vector space.

1.1.22 Exercise. Prove that the set $\mathscr{P} = \{a_0 + a_1x + a_2x^2 + \cdots + a_nx^n\ ;\ n \in \mathbb{N} \text{ and } a_0, ..., a_n \in \mathbb{K}\}$ is a vector space. Prove also that $\mathscr{P}$ is infinite dimensional. Prove that $C((0,1))$ is infinite dimensional.

1.1.23 Exercise. Define the set $C^\infty((0,1)) = \cap_{k=0}^{\infty} C^k((0,1))$. (Notice that $C^\ell((0,1)) \subset C^k((0,1))$ when $\ell \geq k$.) Prove that $C^\infty((0,1))$ is a vector space.

1.1.24 Exercise. Prove that a closed subset of a complete metric space is complete.

1.1.25 Exercise. Let X and Y be metric spaces and $f : X \to Y$ be continuous. Show by counter example that the condition x_n is Cauchy in X need not imply that $y_n = f(x_n)$ is Cauchy in Y.

1.1.26 Exercise. Let X and Y be metric spaces and $f : X \to Y$ be uniformly continuous. Show that $y_n = f(x_n)$ is Cauchy in Y whenever x_n is Cauchy in X.

Assume in addition that X and Y are complete. Show that the limit $x = \lim x_n$ in X satisfies the equation $y = f(x)$ when $y = \lim y_n$ in Y.

1.1.27 Exercise. Apply the result of the previous exercise to prove that the bisection algorithm converges to $\sqrt{a}$ when $a > 1$ and we begin with the interval $0 \le x \le a$. What happens when $a < 1$?

1.1.28 Exercise. (triangle inequality) Let X be a normed vector space. Show that $\|x-y\| \le \|x-z\| + \|z-y\|$ for every $x, y, z \in X$. Show that the function $d(x, y) = \|x - y\|$ is a metric on X, i.e., that it satisfies (a) $d(x, y) \ge 0$ for all $x, y \in X$, (b) $d(x, y) = 0$ implies $x = y$, and (c) $d(x, z) \le d(x, y) + d(y, z)$ for all $x, y, z \in X$.

1.1.29 Exercise. Verify the properties of a norm for the function $\|\cdot\|$ in (1.1.1).

1.1.30 Exercise. Let $[a, b]$ be a closed bounded interval in $\mathbb{R}$ and denote by $C([a, b]; \mathbb{R}^n)$ the set of continuous functions on $[a, b]$ which are valued in $\mathbb{R}^n$. Prove that $C([a, b]; \mathbb{R}^n)$ is a Banach space.

1.1.31 Exercise. Let $[a, b]$ be a closed bounded interval in $\mathbb{R}$ and denote by $C^1([a, b])$ the set of continuous functions on $[a, b]$ which are continuously differentiable on $[a, b]$. (Assume the necessary one-sided limits exist at a and b.) Prove that $C^1([a, b])$ is a Banach space using the norm

$$\|f\|_1 \;=\; \|f\| + \|f'\|$$

where $\|\cdot\|$ is defined in (1.1.1) and f' denotes the derivative.

1.1.32 Exercise. Prove that both $\mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ are Banach spaces, over $\mathbb{R}$ and $\mathbb{C}$ respectively, when the norm is given by

$$\|A\| \;=\; \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2 \,.$$

1.1.33 Exercise. Show that $f_n(x) = x^n/n^2$ is a Cauchy sequence in $C([0, 1])$. For which closed intervals $[a, b] \subset \mathbb{R}$ is f_n Cauchy in $C([a, b])$?

1.1.34 Exercise. Show that $f_n(x) = \frac{\sin nx}{n^\alpha}$ is Cauchy in $C([a, b])$ if $\alpha > 0$, and this for any interval $[a, b]$.

1.1.35 Exercise. If X is a vector space with norm $\|\cdot\|$, a second norm $\|\cdot\|'$ on X is *equivalent* to $\|\cdot\|$ if there are positive constants c and C such that

$$c\,\|x\| \;\le\; \|x\|' \;\le\; C\,\|x\|$$

for all $x \in X$. In this setting, show that there exist two other positive constants c' and C' such that

$$c'\,\|x\|' \;\le\; \|x\| \;\le\; C'\,\|x\|'$$

for all $x \in X$. Let x_n be a sequence in X. Show that $\|x_n - x\| \to 0$ as $n \to \infty$, for some x, if and only if $\|x_n - x\|' \to 0$ as $n \to \infty$. Show that x_n is Cauchy with respect to $\|\cdot\|$ if and only if it is Cauchy with respect to $\|\cdot\|'$. Show that $A \subset X$ is open with respect to $\|\cdot\|$ if and only if it is open with respect to $\|\cdot\|'$. If Y is a metric space and $f : X \to Y$, show that f is continuous with respect to $\|\cdot\|$ if and only if it is continuous with respect to $\|\cdot\|'$.

1.1.36 Exercise. Let Q be an $n \times n$ matrix of complex numbers which is Hermitian, that is, $Q^* = Q$ where Q^* is the complex conjugate transpose of Q. Assume also that Q is (strictly) positive definite, that is, $x^*Qx > 0$ for all non-zero $x \in \mathbb{C}^n$. Show that the function $(x, y) \mapsto x^*Qy$, of the two vector variables x and y in $\mathbb{C}^n$, is an inner product on $\mathbb{C}^n$.

Formulate the corresponding statement for an inner product over $\mathbb{R}^n$ when Q is real and symmetric.

1.1.37 Exercise. Prove the *parallelogram law*

$$\|x + y\|^2 + \|x - y\|^2 = 2\,(\,\|x\|^2 + \|y\|^2\,)$$

when x and y are elements of any inner product space.

1.1.38 Exercise. Consider the Hilbert space $H^1(0,1)$ and the sesqui-linear form $s(u,v) = \int_0^1 u'(x)\bar{v}'(x)\,dx$. Is s an inner product on $H^1(0,1)$? Prove your answer.

Answer the same question with $H_0^1(0,1)$ in place of $H^1(0,1)$. Consider the fact that $u(x) = \int_0^x u'(t)\,dt$.

Define the sesqui-linear form $r(u,v) = u(0)\bar{v}(0) + u(1)\bar{v}(1) + \int_0^1 u'(x)\bar{v}'(x)\,dx$ for $u, v \in H^1(0,1)$. Is r an inner product on $H^1(0,1)$? Prove your answer. (Hint: if $u \in H^1(0,1)$, consider the function $v(x) = u(x) - (u(0)(1-x) + u(1)x)$ which is zero at 0 and 1.)

1.1.39 Exercise. Define the set $V = \{f \in C^\infty([0,1])\ ;\ f(0) = 0\}$. Is V a vector space? Is V an inner product space with inner product $\langle f, g\rangle = \int_0^1 f(x)\bar{g}(x) + f'(x)\bar{g}'(x)\,dx$? Prove your answers. If your answer is 'yes' to both questions, describe the Hilbert space completion of V in terms of the behavior of the functions at $x = 0$ and $x = 1$.

1.1.40 Exercise. Let X and Y be Banach spaces and $f : X \to Y$ be linear and uniformly continuous. Assume that $\sum_1^\infty x_n$ is absolutely convergent in X. Show that $\sum_1^\infty f(x_n)$ is convergent in Y and that $\sum_1^\infty f(x_n) = f(\sum_1^\infty x_n)$.

1.1.41 Exercise. Use trigonometric angle addition formulas and the definition $e^{i\theta} = \cos\theta + i\,\sin\theta$ when $\theta \in \mathbb{R}$ to show that exponents add: $e^{i\alpha}e^{i\beta} = e^{i(\alpha+\beta)}$. Show that $e^{-i\theta} = \cos\theta - i\,\sin\theta$, the complex conjugate of $e^{i\theta}$.

Let $m, n \in \mathbb{N}$. Use a table of integrals to compute the inner product

$$\int_0^\pi \sin(mt)\,e^{-int}\,dt\,.$$

1.1.42 Exercise. Use the definition of the complex exponential in the previous exercise to prove the formulas

$$\sin\theta \;=\; \frac{e^{i\theta} - e^{-i\theta}}{2i} \qquad \text{and} \qquad \cos\theta \;=\; \frac{e^{i\theta} + e^{-i\theta}}{2}\,.$$

Use the formula for sin to compute the inner product

$$\int_0^\pi \sin(mt + \phi)\,e^{-int}\,dt$$

where $\phi \in \mathbb{R}$ is an unknown phase and $m, n \in \mathbb{N}$. (You need to use the fact that an anti-derivative of e^{ict} with respect to t is $e^{ict}/(ic)$ where c is any real or complex constant.)

1.1.43 Exercise. Outline plausible computational steps to check whether an image of a face from a TV monitor is the same person as one of the photos in a database of pictures of faces. Assume each picture or image is an array of pixels, each a grayscale value (from 0 to 1), or three arrays of RGB values. Include steps to match the location of the two faces in their images and to measure the match between key facial features (distance between the eyes, distance from tip of nose to upper lip, length of nose, perhaps color of eyes and hair, etc.).

Who would be interested in buying such technology if it could be matured to the point that it was 99% reliable?

1.1.44 Exercise. Consider the Netflix problem of presenting to the user new films and shows that are appealing to him or her. Make a list of at least six features (genre, stars, running time, etc.) that could be used to match program to viewer. Create an algorithm to compare your features, and that could be used to match viewer preferences to programs.

1.1.45 Exercise. Protecting the US President is the job of the Secret Service (SS). Suppose the SS wants a computer program that will scour millions of internet text documents looking for phrases of derogatory statements about the President.

Make a list of key words and phrases that an automated taxt processor should find for this application. Work out some processing details for two of your phrases of interest, to include identifying whether the writer's sentiment is positive or negative toward the President, and if negative, how strongly negative those feelings are.

Try to distinguish between documents that indicate simple disagreement of political opinion and those for which the author may be inclined to take criminal action.

1.2 Vector Spaces of Functions

The spaces of functions defined in this section, sometimes with minor variations, will be used for the applications treated in these notes. We will begin by discussing some useful functions, sometimes called 'test functions' because they play the role of an independent variable, akin to a test particle or test charge in physics.

Functions with compact support These functions will occasionally be needed as 'test functions' in proofs. Their importance can be illustrated by the following example. If f is a continuous function on an interval $(a,b) \subset \mathbb{R}$, and if

$$\int_a^b f(x)g(x)\,dx = 0 \qquad \text{for every continuous function } g \text{ on } (a,b), \tag{1.2.1}$$

then $f(x) = 0$ for all $x \in (a,b)$. (One way to argue this is to set $g = \overline{f}$; then we know from calculus that $\int_a^b f(x)g(x)\,dx = \int_a^b |f(x)|^2\,dx > 0$ unless $f = 0$.) The point of this set up is that it is often easier to show property (1.2.1) than it is to show directly that $f(x) = 0$ at every $x \in (a,b)$. (In applications f is often a complicated expression, or the difference of two functions the equality of which is to be shown, etc.) Because the existence of the integral (1.2.1) is sometimes suspect, it is often useful to apply such a property when g belongs to a more restricted class of functions. One such restriction is that g be smoother (have more derivatives) than just continuous and another restriction is that g equals zero except in a very localized area of interest. We will explore both these restrictions.

1.2.1 Definition. If $\Omega \subset \mathbb{R}^d$ and $f : \Omega \to \mathbb{K}$, the *support* of f is the closure of $\{x \in \Omega\ ;\ f(x) \neq 0\}$ in Ω. That is, the support of f is the smallest closed set in Ω which contains the set where f is not zero. We will denote the support of f by $\mathrm{spt}(f)$.

If the support of f is bounded in $\mathbb{R}^d$ we say f has *compact support*.

A similar definition holds when X is any vector space and $f : \Omega \to X$:

$$\mathrm{spt}(f) = \overline{\{x \in \Omega\ ;\ f(x) \neq 0 \in X\}}\,.$$

For instance, if $\Omega = \mathbb{R}$ and $f(x) = 0$ when $x \leq 0$ and $= x$ when $x > 0$ then $\mathrm{spt}(f) = [0,\infty)$. If the domain of f is restricted to $(-1,1)$ then $\mathrm{spt}(f) = [0,1)$. If $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}^n$, every component of the vector $f(x)$ is zero when $x \notin \mathrm{spt}(f)$.

1.2.2 Definition. Let $\Omega \subset \mathbb{R}^d$ be open and $k \in \mathbb{N}_0 \cup \{\infty\}$. We denote by $C^k(\Omega)$ the vector space of k times continuously differentiable, $\mathbb{K}$-valued, functions on Ω. We denote by $C_0^k(\Omega)$ the functions in $C^k(\Omega)$ whose support is a compact subset of $\mathbb{R}^d$ which lies inside Ω.

This definition is also used for $\mathbb{K}^n$-valued functions in which case we use the notation $C^k(\Omega;\mathbb{K}^n)$ and $C_0^k(\Omega;\mathbb{K}^n)$.

If $f \in C_0(\Omega)$ the $f = 0$ in a neighborhood of $\partial\Omega$, the boundary of Ω. Note that $C_0^k(\Omega;\mathbb{R}^n)$ is a vector subspace of $C^k(\Omega;\mathbb{R}^n)$ since $\mathrm{spt}(f+g) \subset \mathrm{spt}(f) \cup \mathrm{spt}(g)$.

Functions in $C^k(\Omega)$ are familiar. Functions in $C_0^k(\Omega)$ may be constructed in a few steps from a basic example; we illustrate the process.

To obtain continuous functions with compact support let $0 \leq r < R$. For $t \geq 0$ define $f_0(t) = 1$ when $t \leq r$, $f_0(t) = 0$ when $t \geq R$, and $f_0(t) = (R-t)/(R-r)$ when $r < t < R$ (the straight line connecting $(r,1)$ and $(R,0)$. For $t \in \mathbb{R}$ set $f_1(t) = f_0(|t|)$. Then f_1 is continuous on $\mathbb{R}$, and $= 1$ on $[-r,r]$, and $= 0$ when $|t| \geq R$.

If $[a,b]$ is any interval in $\mathbb{R}$ we may cloose r so that $2r = b-a$, $R = r+\epsilon$, and then center (by a shift of variables) f_1 at the mid-point $(a+b)/2$ to construct a function f in $C_0(\mathbb{R})$ such that $f = 1$ on $[a,b]$ and $f = 0$ on the complement of $[a-\epsilon, b+\epsilon]$.

If $R = \prod_1^d [a_j,b_j]$, approximation of a d-dimensional box-car function, $1_R(x_1,...,x_d) = \prod_1^d 1_{[a_j,b_j]}(x_j)$, may be accomplished by building an approximation f_j to $1_{[a_j,b_j]}(x_j)$ as before, then setting $f(x_1,...,x_d) = \prod_1^d f_j(x_j)$ on $\mathbb{R}^d$.

Beginning with the same f_0 as above, one can also construct a function $f \in C_0(B(x_0,r))$ which satisfies $f(x) = 1$ for all $x \in B(x_0,r/2)$, where $B(x_0,r) = \{x \in \mathbb{R}^d\ ;\ |x-x_0| < r\}$ is the open ball of radius r centered at x_0. (Exercise 1.2.28.)

A C^1_0 approximation is obtained in the same way if, in constructing f_0, we replace the line connecting $(r,1)$ and $(R,0)$ with the (unique) cubic polynomial $p(t) = at^3 + bt^2 + ct + d$ that satisfies the four end conditions $p(r) = 1$ and $p'(r) = p(R) = p'(R) = 0$. These conditions make both f_0 and f_0' continuous on $(0,\infty)$.

The following two lemmas show how well controlled smooth functions can be.

1.2.3 Lemma. *Let $r > 0$ and $B(0,r) = \{x \in \mathbb{R}^d \,;\, |x| < r\}$ be the ball of radius r centered at $0 \in \mathbb{R}^d$. Then there exists a function $f \in C^\infty_0(B(0,r))$ which satisfies $f(x) \geq 0$ for all $x \in \mathbb{R}^d$ and $\int_{\mathbb{R}^d} f(x)\,dx = 1$.*

Proof. The reader is ask to construct such a function in Exercises 1.2.30 and 1.2.31. □

1.2.4 Proposition. *Let $K \subset \mathbb{R}^d$ be compact and $A \subset \mathbb{R}^d$ be closed, and assume $K \cap A = \emptyset$, the empty set. Then there is a $f \in C^\infty_0(\mathbb{R}^d)$ with $0 \leq f(x) \leq 1$ for all $x \in \mathbb{R}^d$, such that $f(x) = 1$ for all $x \in K$ and $f(x) = 0$ for all $x \in A$.*

Proof. Let $d = \operatorname{dist}(K,A) = \inf\{|x-y| \,;\, x \in K, y \in A\}$. We know $d > 0$ because if $d = 0$ there would be two sequences, $x_n \in K$ and $y_n \in A$, such that $\lim |x_n - y_n| = d = 0$. Since K is compact there is a subsequence x_{n_k} such that $\lim x_{n_k} = x_0 \in K$. But $\lim y_{n_k} = x_0$ as well since $|x_0 - y_{n_k}| \leq |x_0 - x_{n_k}| + |x_{n_k} - y_{n_k}|$. Since A is closed, $x_0 \in A$. So $K \cap A$ would not be empty if $d = 0$.

Now let $K' = \{x \in \mathbb{R}^d \,;\, \operatorname{dist}(x,K) < d/3\}$. And let $h_r \in C^\infty_0(B(0,r))$ have the properties of f in the preceeding lemma with $r < d/3$. Since the convolution of two functions, each with compact support, satisfies

$$\operatorname{spt}(u * v) \;\subset\; \operatorname{spt}(u) + \operatorname{spt}(v)\,,$$

the function

$$f(x) \;=\; \int_{\mathbb{R}^d} 1_{K'}(y)\, h_r(x-y)\, dy$$

has the desired properties, where $1_{K'}$ is the indicator function of the set K'. □

Banach spaces related to $C(K)$ If f_n, $n \in \mathbb{N}$, is a sequence of complex-valued functions on a set X, and f is another complex-valued function on X, we say that f_n *converges pointwise* to f if for every $x \in X$ and $\epsilon > 0$ there is a $n_0 \in \mathbb{N}$, depending on both x and ϵ, such that $n \geq n_0$ implies $|f_n(x) - f(x)| < \epsilon$.

We say the sequence f_n *converges uniformly* to f on X if the δ in the definition of pointwise convergence can be chosen independent of x, that is, if the same δ, depending on ϵ, will work for every $x \in X$.

For example, $f_n(x) = x^n$ converges pointwise to $f(x) = 0$ on the open interval $(0,1)$ but does not converge uniformly. This sequence does however converge uniformly on $(0, 1-\epsilon)$ for any small $\epsilon > 0$ (Exercise 1.2.32).

A sequence of functions f_n on a set K converges uniformly to a limit function f if and only if $|f_n - f|_0 \to 0$ where $|\cdot|_0$ is the norm defined in the next example.

1.2.5 Example ($C(K)$)**.** Let $K \subset \mathbb{R}^d$ be closed and bounded (compact). The set $C(K)$ of scalar-valued, continuous functions on K is a normed vector space when given the sup-norm

$$|f|_0 \;=\; \sup_{x \in K} |f(x)| \;=\; \sup\{|f(x)| \,;\, x \in K\}\,.$$

The reader should be able to check the required properties; one must observe that

$$\sup\{|f(x)+g(x)|; x \in K\} \leq \sup\{|f(x)| + |g(x)|; x \in K\} \leq \sup\{|f(x)|; x \in K\} + \sup\{|g(x)|; x \in K\}\,.$$

The space $C(K)$ is also complete, hence a Banach space. Completeness will be stated in the next theorem.

1.2.6 Theorem. *Let $K \subset \mathbb{R}^d$ be closed and bounded (compact), and let f_n be a sequence in $C(K)$ which converges uniformly to a function f on K. Then f is continuous on K. Therefore $f \in C(K)$ and $C(K)$ is complete with respect to the 'sup norm' $|\cdot|_0$.*

We will leave the "$\frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3}$" proof of this theorem as an exercise for the reader to look up in an elementary text on real analysis.

1.2.7 Example ($C(K;\mathbb{K}^n)$)**.** We can also consider the vector space $C(K;\mathbb{K}^n)$ of $\mathbb{K}^n$-valued, continuous functions on K. In this context we will use the norm

$$|f|_0 \;=\; \sup_{x\in K}|f(x)| \;=\; \sup_{x\in K}(|f_1(x)|^2+|f_2(x)|^2+\cdots+|f_n(x)|^2)^{1/2}$$

where $f(x)=(f_1(x),f_2(x),...,f_n(x))$. (We will often assume our vectors are column vectors.) The space $C(K;\mathbb{K}^n)$ is a Banach space and the reader should show this using the fact that $C(K)$ is complete.

It is sometimes necessary to include the derivatives of a function in the measure of its size. This is done in the next example.

1.2.8 Example ($C^1(K)$)**.** Let $K\subset\mathbb{R}^d$ be compact and $C^1(K)$ be the vector space of scalar-valued continuous functions on K all of whose first order derivatives are also continuous on K. On this vector space we use the norm

$$|f|_1 \;=\; \sup_{x\in K}(|f(x)|+|\partial_1 f(x)|+\cdots+|\partial_d f(x)|)\,.$$

The space $C^1(K)$ is also complete, hence a Banach space.

Hilbert spaces related to $L^2(\Omega)$

1.2.9 Definition ($L^2(\Omega)$)**.** Let $\Omega\subset\mathbb{R}^d$ be an open set. Define the norm

$$\|f\|_0=\left(\int_\Omega |f(x)|^2\,dx\right)^{1/2}.$$

on the vector subspace X of $C^\infty(\overline{\Omega})$ for which it is finite. We then define $L^2(\Omega)$ to be the completion of X with respect to the norm $\|\cdot\|_0$.

The subscript 0 in the notation for the norm indicates that no derivatives are included in this norm.

$L^2(\Omega)$ may also be defined as the set of all measurable functions f on Ω such that $\|f\|_0<\infty$, provided that we identify any two such functions which are equal almost everywhere.[3] In this context the integral appearing in the definition of $\|\cdot\|_0$ is the Lebesgue integral.

$L^2(\Omega)$ is a Hilbert space when given the inner product

$$\langle f,g\rangle_0 \;=\; \int_\Omega \overline{f(x)}\,g(x)\,dx.$$

The complex conjugation is superfluous when $\mathbb{K}=\mathbb{R}$. (See Exercise 1.2.22.)

1.2.10 Definition (Sobolev space $H^1(\Omega)$)**.** Let $d\subset\mathbb{R}^d$ be open and denote by $\tilde{H}^1(\Omega)$ the subspace of $C^\infty(\Omega)$ for which the norm

$$\|f\|_1 \;=\; \left(\sum_{|\alpha|\le 1}\int_\Omega |\partial^\alpha f(x)|^2\,dx\right)^{1/2} \;=\; \left(\int_\Omega |f|^2+|\partial_1 f|^2+\cdots+|\partial_d f|^2\,dx\right)^{1/2} \tag{1.2.2}$$

is finite. Then we define $H^1(\Omega)$ to be the completion of $\tilde{H}^1(\Omega)$ with respect to the norm $\|\cdot\|_1$.

When $k\ge 1$, $H^1(\Omega)$ may also be defined as the completion of those $C^k(\Omega)$ functions for which this norm is finite.

A more descriptive definition is to set

$$H^1(\Omega) \;=\; \{f\in L^2(\Omega)\,;\,\partial_j f\in L^2(\Omega) \text{ for all } j=1,2,...,d\}\,.$$

[3] To identify two functions which are equal almost everywhere means to put them into the same equivalence class. Thus, from the Lebesgue theory, elements of $L^2(\Omega)$ are equivalence classes of measurable functions.

This means that $H^1(\Omega)$ functions are those Lebesgue measureable functions which are square integrable, and whose first order derivatives (exist almost everywhere and) are square integrable.

$H^k(\Omega)$ can also be defined for integers $k > 1$ (the reader may speculate on the norm). Using this notational scheme we sometimes write $H^0(\Omega)$ for $L^2(\Omega)$.

$H^1(\Omega)$ is a Hilbert space with inner product

$$\langle f, g\rangle_1 \;=\; \int_\Omega f(x)\overline{g}(x) + \partial_1 f(x)\,\partial_1\overline{g}(x) + \cdots + \partial_d f(x)\,\partial_d\overline{g}(x)\,dx\ .$$

An important subspace of $H^1(\Omega)$ is the following.

1.2.11 Definition (Sobolev space $H^1_0(\Omega)$)**.** The Hilbert space $H^1_0(\Omega)$ is the completion of $C^\infty_0(\Omega)$ with respect to the norm $\|\cdot\|_1$.

If $\Omega = \mathbb{R}^d$, $H^1_0(\Omega) = H^1(\Omega)$. But if Ω is a bounded open subset of $\mathbb{R}^d$ then $H^1_0(\Omega)$ is a proper subspace of $H^1(\Omega)$: $H^1_0(\Omega)$ is the subspace of $H^1(\Omega)$ functions which equal zero on the boundary $\partial\Omega$. For $H^0(\Omega) = L^2(\Omega)$ this distinction is unimportant; for any open $\Omega \subset \mathbb{R}^d$, $L^2(\Omega)$ may be defined as the $\|\cdot\|_0$ completion of $C^\infty_0(\Omega)$. (See Exercise 1.2.24.)

1.2.12 Remark ($C^1(\overline{\Omega};\mathbb{K}^n)$, $H^1(\Omega;\mathbb{K}^n)$, etc.)**.** All of the preceding examples of function spaces can be generalized to spaces of functions taking values in $\mathbb{K}^n$ instead of $\mathbb{K}$. In each instance we replace $|f(x)|$ on $\mathbb{K}$ by $|f(x)| = \sqrt{|f_1(x)|^2 + \cdots + |f_n(x)|^2}$ when $f(x) \in \mathbb{K}^n$. For instance, when $\Omega \subset \mathbb{R}^d$ the inner product in $H^1(\Omega, \mathbb{C}^n)$ is

$$\langle f, g\rangle_1 \;=\; \int_\Omega \sum_{k=1}^{n}\Big(f_k\overline{g}_k + \sum_{j=1}^{d}\partial_j f_k\,\partial_j\overline{g}_k\Big)\,dx$$

where $f(x) = (f_1(x), ..., f_n(x))$.

1.2.13 Remark (Subspaces)**.** If X is a Banach (Hilbert) space and $M \subset X$ is any sub-vector space, then M is a normed vector space (inner product space) using the norm (inner product) for X. If in addition M is a closed set (a *closed subspace*) then it is also a Banach (Hilbert) space since a closed subset of a complete metric space is itself complete.

1.2.14 Example ($\mathscr{B}(X,Y)$)**.** If X and Y are two Banach spaces, the vector space of bounded linear transformations from X to Y, $\mathscr{B}(X,Y)$, will be defined in the next section. It will be given a norm making it a Banach space. If $X = \mathbb{K}^n$ and $Y = \mathbb{K}^m$ this is the vector space of $m \times n$ metrices with components in $\mathbb{K}$.

1.2.15 Exercise. Show by example that it is possible to have two closed sets A and B in $\mathbb{R}^2$ satisfying $A \cap B = \emptyset$ but $\operatorname{dist}(A,B) = \inf\{|x-y|\ ;\ x \in A, y \in B\} = 0$.

1.2.16 Exercise. Pick some of the examples above and check that they are vector spaces. Also, check that the 'norms' given do indeed satisfy the required properties of a norm.

1.2.17 Exercise. Let $K \subset \mathbb{R}^d$ be compact. Show that $C^1(K)$ is a normed vector space. (Verify both the vector spaces properties, and the properties of the norm.)

1.2.18 Exercise. Let $K \subset \mathbb{R}^d$ be compact and $k \in \mathbb{N}$. Generalize the last exercise by showing that $C^k(K)$ is a normed vector space.

1.2.19 Exercise. For $x = (x_1, ..., x_n)$ show that

$$|x| = \sqrt{|x_1|^2 + \cdots + |x_n|^2} \qquad \text{and} \qquad |x|' = |x_1| + \cdots + |x_n|$$

are equivalent norms on $\mathbb{R}^n$.

For functions $f(x) = f(x_1, ..., x_d)$ defined on a compact set $K \subset \mathbb{R}^d$ show that

$$|f|_1 = \sup_{x\in K}\sqrt{|f|^2 + |\partial_1 f|^2 + \cdots + |\partial_d f|^2} \qquad \text{and} \qquad |f|'_1 = \sup_{x\in K}\big(|f| + |\partial_1 f| + \cdots + |\partial_d f|\big)$$

are equaivalent norms on $C^1(K)$.

1.2.20 Exercise. Let X be a Banach space and $K \subset \mathbb{R}^d$ be compact. Define $C(K;X)$ to be the vector space of X-valued continuous functions on K. Show that $C(K;X)$ is a Banach space.

1.2.21 Exercise. Define $C_b((0,1))$ to be the vector space of continuous functions on the open interval $(0,1)$ which are bounded in the norm $|f|_0 = \sup_{0<x<1} |f(x)|$. Show that $C([0,1]) \subset C_b((0,1)) \subset C((0,1))$, and that each containment is proper, i.e., that there are functions in each super set that are not in the subset.

1.2.22 Exercise. Let $L^2(\Omega)$ be as in Definition 1.2.9. Show that $L^2(\Omega)$ is a Hilbert space. (Note: $L^2(\Omega)$ is complete by definition os you need only show that it is a vector space and that $\langle \cdot, \cdot \rangle_0$ is an inner product. One can show that $\|f+g\|_0^2 \le (\|f\|_0 + \|g\|_0)^2$ by using the inner product and Schwarz' inequality.)

1.2.23 Exercise. Prove that $H^k(\Omega)$ is a Hilbert space when Ω is an open subset of $\mathbb{R}^d$.

1.2.24 Exercise. Let n be an integer ≥ 3 and define $f_n(x)$ for $0 \le x \le 1$ to be the piece-wise linear function which is 1 on $[\frac{1}{n}, 1-\frac{1}{n}]$, the straight line joining $(0,0)$ and $(\frac{1}{n}, 1)$ when $0 \le x < \frac{1}{n}$, and the straight line joining $(1-\frac{1}{n}, 1)$ and $(1,0)$ when $1-\frac{1}{n} < x \le 1$. Show that f_n belongs to $L^2(0,1)$ and to $H_0^1(0,1)$. Show that $f_n \to 1$ in $L^2(0,1)$ but not in the $\|\cdot\|_1$ norm, as $n \to \infty$.

The statement after the definition of $H^1(\Omega)$ is that this function space is equal to the completion of those $C^1(\Omega)$ functions f for which $\|f\|_1 < \infty$. Use your imagination to adjust the definition of the functions f_n defined above (round out the corners) to get functions $f_n \in C_0^1(0,1)$ which satisfy (a) $f_n \to 1$ in $L^2(0,1)$, but (b) f_n does not converge to 1 in $H^1(0,1)$. Argue that $1 \in H^1(0,1)$ but $1 \notin H_0^1(0,1)$. Conclude that $H_0^1(0,1)$ is a proper subspace of $H^1(0,1)$.

Make a similar, slightly informal, argument to show that every straight line $f(x) = a + bx$ lies in $H^1(0,1)$ but no straight line except $f(x) = 0$ lies in $H_0^1(0,1)$.

1.2.25 Exercise. Let P be a probability (measure) on a sample space $\mathbb{S}$, and denote $\mathbb{E}(X) = \int_{\mathbb{S}} X(s)\, dP(s)$, the expectation of the random variable $X : \mathbb{S} \to \mathbb{R}$. Prove that the set of all random variables with finite variance is an inner product space. (What is the inner product?) Prove that the set of all random variables with zero expectation is a subspace, and that on this subspace the covariance of two such random variables is their inner product.

1.2.26 Exercise. Let $L^2(\mathbb{R}; e^{-x^2} dx)$ denote the completion of the set $\{f \in C(\mathbb{R}) \ ; \ \int_{-\infty}^{\infty} |f(x)|\, e^{-x^2}\, dx < \infty\}$ with respect to the norm given by the square root of the integral in this definition. Verify that $L^2(\mathbb{R}; e^{-x^2} dx)$ is an inner product space. Show that every polynomial on $\mathbb{R}$ lies in $L^2(\mathbb{R}; e^{-x^2} dx)$.

1.2.27 Exercise. Let f and g belong to $C_0^k(\Omega)$. Show that $\mathrm{spt}(f+g) \subset \mathrm{spt}(f) \cup \mathrm{spt}(g)$.

1.2.28 Exercise. Let $x_0 \in \mathbb{R}^d$ and $r > 0$, and let $B(x_0, r) = \{x \in \mathbb{R}^d \ ; \ |x - x_0| < r\}$ be the open ball of radius r centered at x_0. Construct a function $f \in C_0(B(x_0, r))$ which satisfies $f(x) = 1$ for all $x \in B(x_0, r/2)$.

1.2.29 Exercise. Let $x_0 \in \mathbb{R}^d$ and $r > 0$, and let $B(x_0, r) = \{x \in \mathbb{R}^d \ ; \ |x - x_0| < r\}$ be the open ball of radius r centered at x_0. Construct a function $f \in C_0^1(B(x_0, r))$ which satisfies $f(x) = 1$ for all $x \in B(x_0, r/2)$.

Briefly sketch an algorithm to construct, for any $k \in \mathbb{N}$, a function $f \in C_0^k(B(x_0, r))$ which satisfies $f(x) = 1$ for all $x \in B(x_0, r/2)$.

1.2.30 Exercise. In this exercise we construct an $f \in C_0^\infty(B(0,1))$ which satisfies $f(x) \ge 0$ for all $x \in \mathbb{R}^d$ and $\int_{\mathbb{R}^d} f(x)\, dx = 1$. For $t \in \mathbb{R}$ let $h(t) = e^{1/t}$ when $t < 0$ and $= 0$ when $t \ge 0$. Check that $h \in C^\infty(\mathbb{R})$, e.g., at $t = 0$.

For $x \in \mathbb{R}^d$ define $g(x) = h(|x|^2 - 1/2)$. Show that $g \in C_0^\infty(\mathbb{R}^d)$ and that $\mathrm{spt}(g) \subset B(0,1)$.

Let $c = \int_{\mathbb{R}^d} g(x)\, dx$. Show that $c > 0$. Set $f(x) = g(x)/c$ and show that f has all the properties stated above.

1.2.31 Exercise. Let $x_0 \in \mathbb{R}^d$, $r > 0$, and $B(x_0, r) = \{x \in \mathbb{R}^d \ ; \ |x - x_0| < r\}$. Show that there is a $f \in C_0^\infty(B(x_0, r))$ which satisfies $f(x) \ge 0$ for all $x \in \mathbb{R}^d$ and $\int_{\mathbb{R}^d} f(x)\, dx = 1$.

1.2.32 Exercise. Let $\epsilon \in (0,1)$. Show that $f_n(x) = x^n$ converges uniformly on $0 \le x \le 1 - \epsilon$. Show that f_n does not converge uniformly on $(0,1)$.

1.2.33 Exercise. Prove Theorem 1.2.6.

1.2.34 Exercise. Use Theorem 1.2.6 to prove that the space $C^1(K)$ is complete.

1.2.35 Exercise. Define $C^1(K;\mathbb{K}^n)$ and give a norm that could be used on this space. Use the previous exercise to prove this vector space is complete.

1.2.36 Exercise. For $x \in \mathbb{R}$ define the *convolution* integral of any two functions f and g to be

$$f * g(x) \;=\; \int_{-\infty}^{\infty} f(x-y)g(y)\,dy$$

whenever this integral exists.

Let $a > 0$ and set $f(x) = 1 + ax$ when $-a < x \leq 0$, $= 1 - ax$ when $0 < x < a$, and zero elsewhere. Let $b > 0$ and set $g(x) = 1$ when $-b < x < b$ and zero elsewhere. Compute $f * f$, $g * g$, and $f * g$.

When 'smoothness' is measured by the number of derivatives a function has, how does convolution effect smoothness?

1.2.37 Exercise. With the convolution $f * g$ define in the previous exercise, show that $f * g = g * f$, that is, that

$$\int_{-\infty}^{\infty} f(x-y)g(y)\,dy \;=\; \int_{-\infty}^{\infty} f(y)g(x-y)\,dy\,.$$

(Use a change of variables $z = x - y$ and $dz = -dy$.)

1.3 Linear Transformations

The terms *space*, *family* and *class* are other words which mean *set*. *Mapping*, *transformation* and *operator* are synonymous with *function*.

1.3.1 Definition. Let X and Y be vector spaces over the same scalar field $\mathbb{K}$ and $f : X \to Y$ a function.

(a) f is *linear* if

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$$

for all $\alpha, \beta \in \mathbb{K}$ and $x, y \in X$.

(b) f is a *functional* if $Y = \mathbb{K}$. If f is also linear we speak of a *linear functional* on X.

Assume in addition that X and Y are *normed* linear spaces.

(c) f is *bounded* if there is a constant $M < \infty$ such that $\|f(x)\|_Y \leq M$ for all $x \in X$.

(d) f is *continuous at* $x_0 \in X$ if for all $\epsilon > 0$ there exists a $\delta > 0$ (which may depend on ϵ and x_0) such that $\|x - x_0\|_X < \delta$ implies $\|f(x) - f(x_0)\|_Y < \epsilon$. We say f is *continuous on* X if f is continuous at every point of X.

(e) If f is linear, we say f is *bounded* if there is a constant $M < \infty$ such that $\|f(x)\|_Y \leq M\|x\|_X$ for all $x \in X$.

(f) If f is linear and bounded we define the *operator norm* of f to be the non-negative real number

$$\|f\| \stackrel{def}{=} \sup\{\|f(x)\| \;;\; \|x\| \leq 1\} < \infty.$$

(g) If f is linear we define its null space and range to be the sets

$$\begin{aligned} \mathscr{N}(f) &= \{x \in X \;;\; f(x) = 0\} \\ \mathscr{R}(f) &= \{y \in Y \;;\; y = f(x) \text{ for some } x \in X\}\,. \end{aligned}$$

(h) The set of all bounded linear operators from X to Y is denoted $\mathscr{B}(X,Y)$. When $Y = X$ we usually write $\mathscr{B}(X)$.

1.3.2 Remark (on the definition)**.** (a) The algebraic operations on the left side of this equation are operations in X; on the right side they are operations in Y. As is the custom in linear algebra, we will usually denote the application of a linear function f to x by fx rather than $f(x)$.

Since X and Y are groups under vector space addition, this definition makes f a group homomorphism. We will shortly state some important consequences of this.

(b) When $Y = \mathbb{R}$, non-linear functionals occur in optimization problems. When $Y = \mathbb{R}$ or $\mathbb{C}$, linear functionals are extremely important tools in analysis. For instance they define "coordinates" on X, whether X is finite or infinite dimensional.

(c) This same definition applies even if X has no norm or vector space structure.

(d) This is the same definition as continuity in any metric space.

(e) No linear function, except the identically zero one, can be bounded according to the usual definition (c) of boundedness. So this concept is superfluous for linear functions. The concept for linear functions in (e) is that M is a bound on the maximum slope or directional derivative.

(f) There are three norms involved in this definition. The norms of both X and Y are used; and the "norm" $\|f\|$ of f is in fact also a norm on the vector space $\mathscr{B}(X, Y)$. (This will be shown shortly.) Usually it will be unnecessary to use a different notation for each norm since the context (what is between the bars $\|\cdot\|$) makes each one clear.

(g) The null space is sometimes called the kernel. In general the set $\mathscr{R}(f)$ is called the *image*, and *range* refers to the set Y. Thus the terminology here, for linear functions, flies in the face of the usual definitions; but its use in functional analysis seems entrenched and we will follow the custom.

It is useful to keep in mind the following facts about the *operator norm*.

1.3.3 Proposition. *Let $L : X \to Y$ be a bounded linear operator. Then*

$$\begin{aligned}\|L\| &= \sup\{\|Lx\| \ ; \ \|x\| < 1\} \\ &= \sup\{\|Lx\| \ ; \ \|x\| = 1\} \\ &= \sup\{\|Lx\|/\|x\| \ ; \ x \neq 0\}.\end{aligned}$$

Further, the non-negative real number $\|L\|$ is the smallest one which satisfies

$$\|Lx\| \leq \|L\| \, \|x\|$$

for all $x \in X$.

Proof. We will leave this proof as an exercise in logic and set theory. For instance, equality of the second and third lines can be shown by using

$$\tfrac{1}{\|x\|}\|Lx\| \ = \ \|L(\tfrac{x}{\|x\|})\|$$

where we have used a defining property of a norm and the linearity of L. □

1.3.4 Example. Let $\lambda : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$\lambda x = \begin{pmatrix}\lambda_1 & \lambda_2\end{pmatrix}\begin{pmatrix}x_1 \\ x_2\end{pmatrix} = \lambda_1 x_1 + \lambda_2 x_2.$$

We claim $\|\lambda\| = \sqrt{\lambda_1^2 + \lambda_2^2}$. The fact that $\|\lambda\| \leq \sqrt{\lambda_1^2 + \lambda_2^2}$ follows immediately from the Cauchy-Schwartz inequality on $\mathbb{R}^2$, $|x \cdot y| \leq |x|\,|y|$. For this says $|\lambda x| \leq \sqrt{\lambda_1^2 + \lambda_2^2}\,|x| \leq \sqrt{\lambda_1^2 + \lambda_2^2}$ if $|x| \leq 1$. To show $\|\lambda\| \geq \sqrt{\lambda_1^2 + \lambda_2^2}$ we can exhibit an x of length one such that $\lambda x = \sqrt{\lambda_1^2 + \lambda_2^2}$. Clearly $x = (\lambda_1, \lambda_2)'/\sqrt{\lambda_1^2 + \lambda_2^2}$ is such a vector. (The $'$ here denotes the transpose of the row vector.) The graph of the function λ is a plane through the origin in $\mathbb{R}^3$, of course. The operator norm $\|\lambda\|$ is the maximum slope this plane achieves as we look in any direction away from 0.

1.3.5 Example. Let $A : \mathbb{C}^n \to \mathbb{C}^n$ be Hermitian (that is $A^* = A$). Then A has d real eigenvalues $\lambda_1, ..., \lambda_d$ and a complete orthonormal set of (column) eigenvectors $v_1, ..., v_d$, and if we set $V = [v_1, ..., v_d]$ and $\Lambda = \operatorname{diag}(\lambda_1, ..., \lambda_d)$ then $A = V\Lambda V^*$. It is now not hard to show that $\|A\| = \max_{1 \leq j \leq d} |\lambda_j|$.

1.3.6 Example. In general if $A : \mathbb{C}^n \to \mathbb{C}^m$ is linear, $\|A\| = \max_{1\le j\le d} \sigma_j$ where $\{\sigma_j\}$ is the set of singular values of A, i.e., the eigenvalues of the (positive semi-definite Hermitian) matrix A^*A. For if $A = U\Sigma V^*$ where U and V are unitary, then

$$\begin{aligned}\sup_{\|x\|=1} \|A\| &= \big(\sup_{\|x\|=1} \|A\|^2\big)^{1/2} = \big(\sup x^*V\Sigma^2V^*x\big)^{1/2} = \\ &= \big(\sup_{\|y\|=1} y^*\Sigma^2 y\big)^{1/2} = \big(\max_{1\le i\le n} \sigma_i^2\big)^{1/2} = \max\ \sigma_i\ .\end{aligned}$$

If $L : X \to Y$ is a linear transformation and X is a finite dimensional vector space then L is continuous. It is interesting that there are linear operators on infinite dimensional normed vector spaces X which are *not* continuous.[4]

1.3.7 Example. Let X be the vector space of continuously differentiable functions on a closed interval $[a, b]$, Y be the vector space of continuous functions on $[a, b]$, and $L = \frac{d}{dt} : X \to Y$. Let X and Y both be given the norm

$$\|f\| = \sup_{a\le t\le b} |f(t)|\ .$$

Consider the following sequence of functions:

$$f_n(t) = \sin(nt)/\sqrt{n}\ .$$

When n sufficiently large, $\|f_n\| = 1/\sqrt{n}$, and $\|f_n'\| = \sqrt{n}$. So $f_n \to 0$ in X but $f_n' \not\to 0$ in Y. Thus $L(\lim f_n) = L(0) = 0 \ne \lim Lf_n$ and L is not continuous.

The following theorem gives a striking connection between continuity and boundedness for linear operators.

1.3.8 Theorem. *Let X and Y be normed vector spaces and $L : X \to Y$ be a linear mapping. Then the following are equivalent:*

a) L is continuous on X.

b) L is continuous at a single point $x \in X$.

c) L is bounded.

Proof. As in any metric space, continuity of L is equivalent to sequential continuity: $\lim_{n\to\infty} Lx_n = L(\lim_{n\to\infty} x_n)$.

(a) $\Rightarrow$ (b): This is trivial.

(b) $\Rightarrow$ (a): Assume L is continuous at $z \in X$. Then $Lx \to Lz$ in Y as $x \to z$ in X. But this is the same as saying that $L(x - z) \to 0$ in Y as $x - z \to 0$ in X. Replacing $x - z$ by x in this last statement means that L is continuous at 0.

Now if z' is any point in X, and if $x \to z'$, then $L(x - z') \to 0$ in Y as $x - z' \to 0$ in X since we now know that L is continuous at 0. But this is the same as saying $Lx \to Lz'$ as $x \to z'$; so L is continuous at z'.

(c) $\Rightarrow$ (a): For any x and z in X we have

$$\|Lx - Lz\| = \|L(x - z)\| \le \|L\|\,\|x - z\|\,.$$

Thus, $Lx \to Lz$ if $x \to z$; so L is continuous at z.

(a) $\Rightarrow$ (c): Since L is continuous at 0, for every $\epsilon > 0$ there is a $\delta > 0$ such that $\|x\| < \delta$ implies $\|Lx\| < \epsilon$. Thus $\sup\{\|Lx\|\ ;\ \|x\| < \delta\} \le \epsilon$.

Each of the following expressions now implies the next:

$$\begin{aligned}&\sup\{\|Lx\|\ ;\ \|x\| < \delta\} \le \epsilon\\ &\sup\{\tfrac{1}{\delta}\|Lx\|\ ;\ \|x\| < \delta\} \le \epsilon/\delta\\ &\sup\{\|L(\tfrac{x}{\delta})\|\ ;\ \|x\|/\delta < 1\} \le \epsilon/\delta\\ &\sup\{\|Lx\|\ ;\ \|x\| < 1\} \le \epsilon/\delta\end{aligned}$$

The last step is obtain by replacing the dummy argument x/δ by x; and the last expression means $\|L\| \le \epsilon/\delta$. □

[4] However, in this case X cannot be complete by Banach's Closed Graph Theorem; see Curtain & Pritchard, p 45.

1.3.9 Example. Let $L : C^1([0,1]) \to C([0,1])$ be defined by $Lf(t) = \frac{df}{dt}(t)$. The relevant norms are

$$|f|_1 = \sup_{0\le t\le 1} |f(t)| + \sup_{0\le t\le 1} |f'(t)| \qquad \text{and} \qquad |f|_0 = \sup_{0\le t\le 1} |f(t)| \ .$$

Clearly $\frac{df}{dt} \in C([0,1])$ if $f \in C^1([0,1])$. We claim L is continuous (bounded). For

$$|Lf|_0 = |\frac{df}{dt}|_0 \le |f|_0 + |\frac{df}{dt}|_0 = |f|_1.$$

The inequality shows that $\|L\| \le 1$.

1.3.10 Example. This is a variation of the preceding example. Let $L : H^1((0,1)) \to H^0((0,1)) = L^2((a,b))$ be defined by $Lf(t) = \frac{df}{dt}(t)$. The relevant norms are

$$\|f\|_1 = \Big(\int_a^b |f(t)|^2 + |f'(t)|^2 \, dt\Big)^{1/2} \qquad \text{and} \qquad \|f\|_0 = \Big(\int_a^b |f(t)|^2 \, dt\Big)^{1/2} \ .$$

Clearly $\frac{df}{dt} \in H^0((0,1))$ if $f \in H^1((0,1))$. And L satisfies

$$\|Lf\|_0^2 \ = \ \int_0^1 |\frac{df}{dt}|^2 \, dt \ \le \ \int_0^1 |f|^2 + |\frac{df}{dt}|^2 \, dt = \ \|f\|_1^2.$$

So $\|L\| \le 1$.

1.3.11 Example. Let $\Omega \subset \mathbb{R}^d$ be a bounded open or closed set and $L : C(\overline{\Omega}) \to C(\overline{\Omega})$ be defined by

$$Lf(x) = \int_\Omega k(x,y) f(y) \, dy$$

where $k \in C(\overline{\Omega} \times \overline{\Omega})$. L is bounded because

$$\begin{aligned} |Lf|_0 &= \sup_{x\in\Omega} |\int_\Omega k(x,y) f(y) \, dy| \le \sup_{x\in\Omega} \int_\Omega |k(x,y) f(y)| \, dy \\ &\le \int_\Omega \sup_{x\in\Omega} |k(x,y)| \, |f(y)| \, dy \le \int_\Omega \sup_{x,z\in\Omega} |k(x,z)| \, |f(y)| \, dy \\ &\le \int_\Omega \sup_{x,z\in\Omega} |k(x,z)| \, \sup_{y\in\Omega} |f(y)| \, dy \le \mathrm{vol}(\Omega) \, |k|_0 \, |f|_0 \end{aligned}$$

where $|k|_0$ is the sup norm of k over the set $\overline{\Omega} \times \overline{\Omega}$. This bound shows that $\|Lf\| \le \mathrm{vol}(\Omega) \, |k|_0$. It also shows that $Lf \in C(\overline{\Omega})$ if $f \in C(\overline{\Omega})$ so the image of L genuinely lies in $C(\overline{\Omega})$.

1.3.12 Example. Let $L : L^2((-1,1)) \to \mathbb{C}$ be given by $Lf = f(0)$. The L^2 norm is $\|f\| = (\int_{-1}^1 |f(t)|^2 \, dt)^{1/2}$. This L is not continuous. For $n \in \mathbb{N}$ set

$$f_n(t) \ = \ \begin{cases} 0 & \text{for } t < -\frac{1}{n}, \\ 1 + nt & \text{for } -\frac{1}{n} \le t < 0, \\ 1 - nt & \text{for } 0 \le t < \frac{1}{n}, \\ 0 & \text{for } t > \frac{1}{n}. \end{cases} \tag{1.3.1}$$

We see that $f_n(0) = 1$ for all n. But $\int_{-1}^1 |f_n(t)|^2 \, dt \le \int_{-1}^1 |f_n(t)| \, dt \to 0$ as $n \to \infty$. Therefore $\lim_n f_n(t) = 0$ in $L^2((-1,1))$, and $\lim_n L(f_n) = 1 \ne 0 = L(\lim_n f_n)$.

1.3.13 Example. Let $L : C([-1,1]) \to \mathbb{C}$ be given by $Lf = f(0)$. Let $|f|_0 = \sup_{-1\le t\le 1} |f(t)|$ be the norm on $C([-1,1])$. Then L is continuous since $|f(0) - f_n(0)| \le |f - f_n|_0$.

It is sometimes convevient to define a linear transformation on a dense subset of a Banach space but difficult to define it on the whole space. When the linear transformation is continuous the following result can be used to extend its domain to the whole space.

1.3.14 Proposition. *Let X and Y be Banach spaces, $X_0 \subset X$ a subspace which is dense in X, and let $L_0 : X_0 \to Y$ be a bounded linear transformation. Then there is a unique bounded linear transformation $L : X \to Y$ such that $L = L_0$ on X_0. Further, $\|L\| = \|L_0\|$.*

Proof. We first show how to define L. If $x \in X_0$ we set $Lx = L_0 x$ of course. Suppose $x \notin X_0$. Then there is a sequence x_n in X_0 such that $x_n \to x$ as $n \to \infty$. For notational convenience set $y_n = L_0 x_n$. The bound

$$\|y_n - y_m\| \;=\; \|L_0(x_n - x_m)\| \;\le\; \|L_0\|\,\|x_n - x_m\|$$

shows that y_n is Cauchy in Y. (For since x_n is Cauchy, if $\epsilon > 0$ is given we have $\|x_n - x_m\| < \epsilon/\|L_0\|$ if m and n are sufficiently large; but then $\|y_n - y_m\| < \epsilon$ for this same m and n.) Since Y is complete there is a $y \in Y$ such that $y_n \to y$. Now set $Lx = y$.

Next we check that L is well-defined. Suppose $\tilde{x}_n$ is another sequence in X_0 converging to x. Set $\tilde{y}_n = L_0\,\tilde{x}_n$ and $\tilde{y} = \lim\, \tilde{y}_n$. L is well defined if $y = \tilde{y}$. But

$$\|y_n - \tilde{y}_n\| \;=\; \|L_0\,(x_n - \tilde{x}_n)\| \;\le\; \|L_0\|\,\|x_n - \tilde{x}_n\|\,.$$

Since $x_n - \tilde{x}_n \to 0$ as $n \to \infty$ we also have $y_n - \tilde{y}_n \to 0$, and therefore $y = \tilde{y}$.

We check that L is linear. Let $x, z \in X$ and $\alpha, \beta \in \mathbb{K}$, and let x_n and z_n be sequences in X_0 such that $x_n \to x$, $z_n \to z$, and $\alpha x_n + \beta z_n \to \alpha x + \beta z$. Using the definition of L is the first and third equalities, and the linearity of L_0 for the second, we have

$$L(\alpha x + \beta z) \;=\; \lim\, L_0(\alpha x_n + \beta z_n) \;=\; \lim\,(\alpha\, L_0\, x_n + \beta\, L_0\, z_n) \;=\; \alpha\, Lx + \beta\, Lz\,;$$

so L is linear.

Finally we show that $\|L\| = \|L_0\|$. We have

$$\|L\| \;=\; \sup\{\|Lx\|\;;\; x \in X \text{ and } \|x\| \le 1\} \;=\; \sup\{\|Lx\|\;;\; x \in X_0 \text{ and } \|x\| \le 1\} \;=\; \sup\{\|L_0 x\|\;;\; x \in X_0 \text{ and } \|x\| \le 1\} \;=\; \|L_0\|\,.$$

The second equality holds because X_0 is dense in $\{x \in X\;;\; \|x\| \le 1\}$. □

1.3.15 Example. In Example 1.2.11 we defined $H^1(\Omega)$ as the completion of $C_b^\infty(\Omega)$ with respect to the norm (1.2.2) with $k = 1$ and $p = 2$. If $\Omega = (0,1)$ it is clear what $\frac{d}{dx} f(x)$ means when $f \in C_b^\infty(\Omega)$. The preceding proposition shows that these values alone are sufficient to define the bounded linear transformation $\frac{d}{dx}$ on $H^1((0,1))$ (cf. Example 1.3.10). For instance the function

$$f(t) \;=\; \begin{cases} 0 & \text{for } 0 < t \le \frac{1}{2}, \\ 2t - 1 & \text{for } \frac{1}{2} < t < 1 \end{cases} \tag{1.3.2}$$

is in $H^1((0,1))$ but is not C^1 on $(0,1)$. How do we 'define' $\frac{d}{dx} f(1/2)$? The answer is that we may not need to define a number $\frac{d}{dx} f(x)$ for every $x \in (0,1)$. The function $\frac{d}{dx} f$ only has to exist as an element of $L^2((0,1)) = H^0((0,1))$, and these functions may not have specific values at *every* $x \in (0,1)$ (cf. Example 1.3.12).

1.3.16 Theorem. *If X and Y are normed linear spaces then $\mathscr{B}(X,Y)$ is a normed vector space when given the operator norem. The operator norm satisfies*

$$\|L_1 + L_2\| \;\le\; \|L_1\| + \|L_2\| \qquad \textit{and} \qquad \|\alpha L\| \;=\; |\alpha|\,\|L\| \tag{1.3.3}$$

when L_1, L_2, and L are in $\mathscr{B}(X,Y)$ and $\alpha \in \mathbb{K}$. If Y is also a Banach space then $\mathscr{B}(X,Y)$ is a Banach space.

Proof. We leave it as an exercise for the reader to show that (1.3.3) holds and that $\mathscr{B}(X,Y)$ is a normed vector space. We show that $\mathscr{B}(X,Y)$ is complete with the operator norm $\|\cdot\|$.

Let's first define the limit $L = \lim_{k\to\infty} L_k$ for a Cauchy sequence L_k in $\mathscr{B}(X,Y)$. If $x = 0$ in X set $Lx = 0$. If $x \in X$ but $x \neq 0$ let $y_k = L_k x$; we want to define $Lx = \lim y_k$. Since Y is complete this limit exists if y_k is Cauchy. Write

$$\|y_k - y_\ell\| \;=\; \|L_k x - L_\ell x\| \;\leq\; \|L_k - L_\ell\|\,\|x\|\,. \tag{1.3.4}$$

Since L_k is Cauchy, if $\epsilon > 0$ is given and $x \neq 0$ is fixed there is an $n \in \mathbb{N}$ such that $k, \ell \geq n$ implies $\|L_k - L_\ell\| < \epsilon/\|x\|$. The estimate (1.3.4) then shows that y_k is Cauchy. Since $\|Lx - L_k x\| \leq \|Lx - y_\ell\| + \|y_\ell - y_k\|$, we also see that $\|L - L_k\| \to 0$ as $k \to \infty$.

To check that the definition of L is independent of the sequence L_k, let $\tilde{L}_k$ be another sequence in $\mathscr{B}(X,Y)$ for which $L = \lim_{k\to\infty} \tilde{L}_k$. Then for any $x \in X$, $\|Lx - \tilde{L}_k x\| \leq \|L - \tilde{L}_k\|\,\|x\| \to 0$ as $k \to \infty$. So $Lx = \lim_k \tilde{L}_k x$ in Y holds for the sequence $\tilde{L}_k$ as well.

To show that L is linear write $L(\alpha x + \beta z) = \lim L_k(\alpha x + \beta z) = \alpha \lim_k L_k x + \beta \lim_k L_k z$.

To see that L is bounded let $\epsilon > 0$ be given and choose $n \in \mathbb{N}$ so large that $k \geq n$ implies $\|L - L_k\| \leq \epsilon$. Then

$$\|Lx\| \leq \|L_k x\| + \|Lx - L_k x\| \leq \|L_k x\| + \epsilon\,\|x\| \leq (\|L_k\| + \epsilon)\,\|x\|\,.$$

So $\|L\| \leq \|L_k\| + \epsilon$. □

1.3.17 Proposition. *Let X and Y be normed vector spaces and $L : X \to Y$ a linear transformation. Then*

(a) The null space, $\mathscr{N}(L) = \{x \in X; Lx = 0\}$, is a normed vector subspace of X.

(b) The range space, $\mathscr{R}(L) = \{y \in Y\ ;\ y = Lx \text{ for some } x \in X\}$, is a normed vector subspace of Y.

(c) If L is continuous $\mathscr{N}(L)$ is closed in X.

Proof. (a) Exercise: Show that $\mathscr{N}(L)$ is a vector subspace of X, and that the norm it inherits from X is still a norm for $\mathscr{N}(L)$.

(b) Exercise: Show that $\mathscr{R}(L)$ is a vector subspace of Y, and that the norm it inherits from Y is still a norm for $\mathscr{R}(L)$.

(c) $\mathscr{N}(L)$ is closed since $\{0\} \subset Y$ is a closed set, and therefore $\mathscr{N}(L) = L^{-1}(0)$ is also closed by the continuity of L. □

If X is a Banach space, $x \in X$ and $r > 0$, set $B(x,r) = \{y \in X; \|y - x\| < r\}$ the open ball of radius r about x. If the space X is ambiguous, we will write $B_X(x,r)$.

The folowing example illustrates some of the situations that can arise with linear transformations on infinite dimensional vector spaces.

1.3.18 Example. Denote by ℓ^2 the set of all sequences x_n of complex numbers such that $\sum_{n=1}^{\infty} |x_n|^2 < \infty$. For notational ease we write $\mathbf{x} = (x_1, x_2, ...)$ for the entire sequence. Give to ℓ^2 a vector space structure by adding sequences point-wise, and multiplying all components by the same scalar, i.e., $\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, ...)$ and $\alpha\mathbf{x} = (\alpha x_1, \alpha x_2, ...)$. Define an inner product on ℓ^2 by $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{n=1}^{\infty} \overline{x}_n y_n$, and the resulting norm $\|\mathbf{x}\| = (\sum_{n=1}^{\infty} |x_n|^2)^{1/2}$. This structure makes ℓ^2 a Hilbert space (Exercise 1.3.23).

Now let $\{t_n\}_{n=1}^{\infty}$ be any sequence of complex numbers and let $D \subset \ell^2$ be the subset of elements $\mathbf{x}$ of ℓ^2 for which $\sum_1^{\infty} |t_n|^2 |x_n|^2 < \infty$. The set D is a vector subspace, and carries the norm $\|\cdot\|$ inherited from ℓ^2. Define the linear transformation $T : D \to \ell^2$ by

$$T\mathbf{x} = (t_1 x_1, t_2 x_2, t_3 x_3, ...)\,.$$

The subspace D is the *domain* of T. We think of $\mathbf{x}$ as a column vector and T as an infinite dimensional diagonal matrix whose diagonal components are the numbers t_n. We also define the range of T by $R = \{\mathbf{y} \in \ell^2\ ;\ \mathbf{y} = T\mathbf{x} \text{ for some } \mathbf{x} \in \ell^2\}$.

We claim the following statements hold.

a) $D = \ell^2$ if and only if the sequence $\{t_n\}_{n=1}^{\infty}$ is bounded in $\mathbb{C}$.

b) T is bounded if and only if the sequence $\{t_n\}_{n=1}^{\infty}$ is bounded in $\mathbb{C}$.

c) T is one-to-one on D if and only if $t_n \neq 0$ for all $n \in \mathbb{N}$.

d) $T^{-1} : R \to D$ exists if and only if T is one-to-one on D.

e) $R = \ell^2$ if and only if there exists an $\epsilon > 0$ such that $|t_n| \geq \epsilon$ for all $n \in \mathbb{N}$.

f) T^{-1} exists as a bounded linear operator if and only if there is an $\epsilon > 0$ such that $|t_n| \geq \epsilon$ for all $n \in \mathbb{N}$.

g) $T(B) = \{\mathbf{y} \in R\,;\ \mathbf{y} = T\mathbf{x} \text{ for some } \mathbf{x} \in B\}$ has compact closure in ℓ^2 for every bounded set $B \subset \ell^2(\mathbb{N})$ if and only if $t_n \to 0$ in $\mathbb{C}$ as $n \to \infty$.

Before showing these claims we make a few remarks.

To say that a set B is bounded means that there is an $M > 0$ such that $\|\mathbf{x}\| \leq M$ for all $\mathbf{x} \in B$. To say that $T(B)$ has compact closure means any one of the following equivalent conditions: (i) every sequence of points in the closure of $T(B)$ has a convergent subsequence with limit in this closure, (ii) every sequence in $T(B)$ has a convergent subsequence with limit in the closure of $T(B)$, or (iii) every sequence in $T(B)$ has a convergent subsequence with limit in ℓ^2.

It is not true that 'closed and bounded' is equivalent to 'compact' in the infinite dimensional vector space ℓ^2. These properties are only equivalent in finite dimensional vector spaces. For instance, the sequence of 'natural basis vectors' $\mathbf{e}_n$, consisting of all zeros except for a 1 in the n-th coordinate, in the closed and bounded set $B = \{\mathbf{x}; \|\mathbf{x}\| \leq 1\}$ has no limit point in ℓ^2. For the squared distance between any two distinct $\mathbf{e}_m$ and $\mathbf{e}_n$ is always $\|\mathbf{e}_m - \mathbf{e}_n\|^2 = \|(..., 0, 0, 1, 0, 0, ..., 0, 0, -1, 0, 0, ...)\|^2 = 1 + 1 = 2$.

Whatever the linear transformation T, i.e., whatever the sequence t_n, each natural basis vector $\mathbf{e}_n$ belongs to the domain D of T. In fact, every finite sequence x (all components are zero except a finite number) belongs to D.

It is a fundamental property of any linear transformation T (and indeed of any group homomorphism) that T is one-to-one if and only if the null space of T is $\mathbf{0}$. To see this observe that $T\mathbf{x} = T\mathbf{y}$ implies $\mathbf{x} = \mathbf{y}$, if and only if $T(\mathbf{x} - \mathbf{y}) = \mathbf{0}$ implies $\mathbf{x} - \mathbf{y} = \mathbf{0}$, if and only if $T\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$.

Operators that satisfy condition (g) are called *compact*; these will be studied in section 1.9. In applications, differential operators often satisfy the condition $t_n \to \pm\infty$ as $n \to \infty$ when a certain basis is used. (Example: $d^2 \sin(nx)/dx^2 = -n^2 \sin(nx)$.) Then T^{-1} is compact.

We now demonstrate the preceding claims.

Proof of the claims. (a) $\Rightarrow$) If t_n is not bounded there is a subsequence, t_{n_k} such that $|t_{n_k}| \geq k$, $k = 1, 2, ...$ Define a sequence of complex numbers x_n by: $x_n = 0$ if $n \neq n_k$ for some k, and $x_{n_k} = 1/k$. We claim $(x_n) \in \ell^2$. For $\sum_{n=1}^{\infty} |x_n|^2 = \sum_{k=1}^{\infty} |x_{n_k}|^2 = \sum_{k=1}^{\infty} 1/k^2 < \infty$. On the other hand, $\sum_{n=1}^{\infty} |t_n|^2 |x_n|^2 = \sum_{k=1}^{\infty} |t_{n_k}|^2 |x_{n_k}|^2 = \sum_{k=1}^{\infty} |t_{n_k}|^2/k^2 \geq \sum_{k=1}^{\infty} 1 = \infty$. So $\mathbf{x} = (x_1, x_2, ...) \notin D$, and we conclude that t_n must be bounded.

$\Leftarrow$) If $|t_n| \leq M$ for all n and $\mathbf{x} = (x_1, x_2, ...) \in \ell^2$ then $\sum_{n=1}^{\infty} |t_n|^2 |x_n|^2 \leq M^2 \sum_{n=1}^{\infty} |x_n|^2 < \infty$. So $\mathbf{x} \in D$.

(b) $\Rightarrow$) If $\{t_n\}$ is unbounded we can pick a subsequence $t_{n_k} \to \infty$ as $k \to \infty$. Then $\|T\mathbf{e}_{n_k}\|/\|\mathbf{e}_{n_k}\| = |t_{n_k}|/1 \to \infty$ as $k \to \infty$. So $\sup\{\|T\mathbf{x}\|\ ;\ \|\mathbf{x}\| = 1\}$ is not finite and T is not bounded.

$\Leftarrow$) If $\|T\| < \infty$ then $|t_n| = \|T\mathbf{e}_n\|/\|\mathbf{e}_n\| \leq \|T\|$ is a bound for every $n \in \mathbb{N}$.

(c) $\Rightarrow$) If any $t_n = 0$ then $T\mathbf{e}_n = \mathbf{0}$ and the null space of T in D includes $\mathbf{e}_n$.

$\Leftarrow$) If all $t_n \neq 0$ and $\mathbf{x} = \sum_n a_n \mathbf{e}_n$ is in D, then $T\mathbf{x} = \mathbf{0}$ implies that $t_n a_n = 0$ for all n. So $a_n = 0$ for all n and $\mathbf{x} = \mathbf{0}$.

(d) $\Rightarrow$) Let $\mathbf{x} \in D$ and $T\mathbf{x} = \mathbf{0}$. Then $\mathbf{x} = T^{-1}(T\mathbf{x}) = T^{-1}(\mathbf{0}) = \mathbf{0}$, so T is one-to-one.

$\Leftarrow$) Let $\mathbf{y} \in R$ be given, and suppose both $\mathbf{x}_1$ and $\mathbf{x}_2$ satisfy $T\mathbf{x}_1 = T\mathbf{x}_2 = \mathbf{y}$. Then $T(\mathbf{x}_1 - \mathbf{x}_2) = \mathbf{0}$ and, since T is one-to-one, we also have $\mathbf{x}_1 - \mathbf{x}_2 = \mathbf{0}$. So for any $\mathbf{y} \in R$ the solution $\mathbf{x} \in D$ of the equation $T\mathbf{x} = \mathbf{y}$ is unique, and T^{-1} is then well defined.

(e) $\Rightarrow$) For every n, $\mathbf{e}_n \in R$ and there exists some $\mathbf{x}_n \in D$ such that $T\mathbf{x}_n = \mathbf{e}_n$. Setting $\mathbf{x}_n = \sum_{m=1}^{\infty} a_{nm}\, \mathbf{e}_m$ we see in particular that $t_n\, a_{nn} = 1$ for every n. Thus neither t_n nor a_{nn} can be zero. By (c) T must be one-to-one, and by (d) $T^{-1} : R \to D$ is algebraically well defined. If no such $\epsilon > 0$ exists, we will show how to construct a $\mathbf{y} \in \ell^2$ with $\mathbf{y} \notin R$; this contradiction will prove the claim.

If $t_n = 0$ for some n the conclusion is obvious, so assume $t_n \neq 0$ for all n. Then if no such ϵ exists there is a subsequence t_{n_k} satisfying $|t_{n_k}| \leq 1/k$, $k = 1, 2, ...$ Now define $\mathbf{y} = \sum_1^{\infty} y_n\, \mathbf{e}_n$ by $y_n = 0$ if $n \notin \{n_k\}_{k=1}^{\infty}$, and

$y_{n_k} = 1/k$. We have $\mathbf{y} \in \ell^2$ because $\|\mathbf{y}\|^2 = \sum_1^\infty (1/k)^2 < \infty$. On the other hand, $T^{-1}\mathbf{y} = \sum_{k=1}^\infty (y_{n_k}/t_{n_k})\,\mathbf{e}_{n_k}$. But $\sum_{k=1}^\infty |y_{n_k}/t_{n_k}|^2 = \sum_{k=1}^\infty |\frac{1/k}{t_{n_k}}|^2 \geq \sum_1^\infty 1^2$ is not finite, so $T^{-1}\mathbf{y} \notin \ell^2$ and $\mathbf{y} \notin R$.

$\Leftarrow$) Assume $|t_n| \geq \epsilon$ and that $\mathbf{y} = (y_1, y_2, ...) \in \ell^2$. For each n define $x_n = y_n/t_n$. Clearly $|x_n| \leq |y_n|/\epsilon$. So $\mathbf{x} = (x_1, x_2, ...) \in \ell^2$, and $T\mathbf{x} = \mathbf{y}$ is also in ℓ^2. Thus T is onto.

(f) $\Rightarrow$) Since T^{-1} is bounded there is a constant $M > 0$ such that $\sup_{\|\mathbf{y}\|=1} \|T^{-1}\mathbf{y}\| \leq M$. In particular taking $\mathbf{y} = \mathbf{e}_n$ shows that $|1/t_n| = \|T^{-1}\mathbf{e}_n\| \leq M$, or that $|t_n| \geq M$.

$\Leftarrow$) If $|t_n| \geq \epsilon > 0$ for all n we know that $R = \ell^2$ from (e), and that T is one-to-one from (c). So $T^{-1} : \ell^2 \to D \subset \ell^2$ is well defined. To show that T^{-1} is bounded write $\|T^{-1}\mathbf{x}\|^2 = \sum_1^\infty |x_n|^2/|t_n|^2 \leq (1/\epsilon^2)\sum_1^\infty |x_n|^2 = (1/\epsilon^2)\|\mathbf{x}\|^2$. So $\|T^{-1}\| \leq 1/\epsilon$.

(g) $\Rightarrow$) Let $B = \{\mathbf{x} \in \ell^2 \ ; \ \|\mathbf{x}\| \leq 1\}$. We assume $T(B)$ has compact closure so if $\mathbf{x}_n$ is any sequence in B, the sequence $T\mathbf{x}_n$ in $T(B)$ will have a convergent subsequence. Now assume, to get a contradiction, that $t_n \not\to 0$. Then there is an $\epsilon > 0$ and a subsequence t_{n_k} which satisfies $|t_{n_k}| \geq \epsilon$ for all k. Consider the infinite sequence $\mathbf{e}_{n_k}$ in B, and its image $T\mathbf{e}_{n_k} = t_{n_k}\,\mathbf{e}_{n_k}$ in $T(B)$. For $k \in \mathbb{N}$ the set of vectors $T\mathbf{e}_{n_k}$ in ℓ^2 are orthogonal, so when $k \neq \ell$ the expression

$$\|T\mathbf{e}_{n_k} - T\mathbf{e}_{n_\ell}\|^2 = \langle T\mathbf{e}_{n_k} - T\mathbf{e}_{n_\ell}, T\mathbf{e}_{n_k} - T\mathbf{e}_{n_\ell}\rangle = |t_{n_k}|^2 + |t_{n_\ell}|^2 \geq 2\,\epsilon^2$$

shows that no point $T\mathbf{e}_{n_k}$ is ever closer than a distance $\sqrt{2}\epsilon$ to any other point in this sequence. So the sequence $T\mathbf{e}_{n_k}$ is not Cauchy and has no convergent subsequence. This contradiction proves the assertion.

$\Leftarrow$) Assume $t_n \to 0$ as $n \to \infty$ and let $B \subset \ell^2$ be any bounded set. We show that every sequence in $T(B)$ has a convergent subsequence. (The limit need not be in $T(B)$.) Let $\mathbf{y}_n = T\mathbf{x}_n$ be such a sequence, with $\mathbf{x}_n \in B$. We will use the notation $\mathbf{y}_n = \sum_{k=1}^\infty b_{nk}\,\mathbf{e}_k$ and $\mathbf{x}_n = \sum_{k=1}^\infty a_{nk}\,\mathbf{e}_k$.

We first show that $\mathbf{y}_n$ has a Cauchy subsequence, that is, that there is a subsequence $\mathbf{y}_{n_k}$ such that for every $\epsilon > 0$ there is an $n_0 \in \mathbb{N}$ such that $k, \ell \geq n_0$ implies $\|\mathbf{y}_{n_k} - \mathbf{y}_{n_\ell}\| \leq \epsilon$. Take $M > 0$ so large that $B \subset \{\mathbf{x} \in \ell^2 \ ; \ \|\mathbf{x}\| \leq M\}$. Then choose $n_0 \in \mathbb{N}$ so large that $n \geq n_0$ implies $|t_n| \leq \epsilon/(4M)$. Decompose $\mathbf{y}_n = \mathbf{y}_n^{(1)} + \mathbf{y}_n^{(2)}$ where the first n_0 components of $\mathbf{y}_n$ are contained in $\mathbf{y}_n^{(1)} = \sum_{m=1}^{n_0} b_{nm}\,\mathbf{e}_m$, and the remaining components in $\mathbf{y}_n^{(2)} = \sum_{m=n_0+1}^\infty b_{nm}\,\mathbf{e}_m$. Then we have

$$\|\mathbf{y}_n - \mathbf{y}_m\|^2 = \|\mathbf{y}_n^{(1)} - \mathbf{y}_m^{(1)}\|^2 + \|\mathbf{y}_n^{(2)} - \mathbf{y}_m^{(2)}\|^2 = \|\mathbf{y}_n^{(1)} - \mathbf{y}_m^{(1)}\|^2 + \sum_{k=n_0+1}^\infty |t_k|^2\,|(a_{nk} - a_{mk})|^2 \ . \tag{1.3.5}$$

For all k, $|(a_{nk} - a_{mk})|^2 \leq 2(|a_{nk}|^2 + |a_{mk}|^2)$, and thus

$$\sum_{k=n_0+1}^\infty |t_k|^2\,|(a_{nk} - a_{mk})|^2 \leq \frac{\epsilon^2}{16M^2}\,2\sum_{k=n_0+1}^\infty (|a_{nk}|^2 + |a_{mk}|^2) \leq \frac{\epsilon^2}{16M^2}\,4M^2 = \frac{\epsilon^2}{4} \tag{1.3.6}$$

for every $\mathbf{y}_m$ and $\mathbf{y}_n$.

Next, each $\mathbf{y}_n^{(1)}$ lies in the closed, bounded set $T(B) \cap \{\mathbf{y} = \sum_1^\infty b_k\,\mathbf{e}_k \in \ell^2 \ ; \ b_k = 0 \text{ for all } k > n_0\}$. This is a subset of the finite dimensional vector space $\mathbb{C}^{n_0}$, and is thus compact. So there is a subsequence $\mathbf{y}_{n_k}$ and an $N \in \mathbb{N}$ such that $k, \ell \geq N$ implies

$$\|\mathbf{y}_{n_k}^{(1)} - \mathbf{y}_{n_\ell}^{(1)}\|^2 \leq \epsilon^2/4$$

. This bound combined with (1.3.5) and (1.3.6) shows that the subsequence $\mathbf{y}_{n_k}$ is Cauchy in $T(B)$. □

If X, Y and Z are a normed linear spaces, and $L : X \to Y$ and $K : Y \to Z$ are both bounded linear transformations, then $K \circ L : X \to Z$ belongs to $\mathscr{B}(X, Z)$. (From now on this composition will be written KL as we do in finite dimensions.) For

$$\|KLx\| \leq \|K\|\,\|Lx\| \leq \|K\|\,\|L\|\,\|x\|$$

for every $x \in X$. $\|K\|$ is the smallest real number for which the first inequality holds, and $\|L\|$ is the smallest real number which can be used in the second. Thus

$$\|KL\| \leq \|K\|\,\|L\|. \tag{1.3.7}$$

The following theorem is very useful because it allows us to exchange the (hard) problem of inverting certain linear transformations to the (easier) problems of raising to powers and taking a limit.

1.3.19 Theorem (geometric series in $\mathscr{B}(X)$)**.** *Let X be a Banach space and $A \in \mathscr{B}(X)$ satisfy $\|A\| < 1$. Then $I - A$ is an invertible (non-singular) linear operator on X and its inverse is given by the geometric series*

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k \tag{1.3.8}$$

which converges in $\mathscr{B}(X)$. Furthermore

$$\|(I - A)^{-1}\| \leq \frac{1}{1 - \|A\|} \qquad \textit{and} \qquad \|(I - A)^{-1} - I + A\| \leq \frac{\|A\|^2}{1 - \|A\|} \ . \tag{1.3.9}$$

Proof. Since $\|A\| < 1$ the geometric series converges:

$$\|\sum_{k=0}^{\infty} A^k\| \ \leq \ \sum_{k=0}^{\infty} \|A^k\| \ \leq \ \sum_{k=0}^{\infty} \|A\|^k \ = \ \frac{1}{1 - \|A\|} \ < \ \infty \ . \tag{1.3.10}$$

(Here the continuity of the norm on $\mathscr{B}(X)$ has been used to pull the limit outside the norm.) Thus, the right side of (1.3.8) belongs to $\mathscr{B}(X)$.

Now the same algebra as in the scalar case shows

$$(I - A)(I + A + A^2 + A^3 + ... + A^n) = (I + A + A^2 + A^3 + ... + A^n)(I - A) = I - A^{n+1}.$$

Since $\|A^{n+1}\| \leq \|A\|^{n+1} \to 0$ as $n \to \infty$, passing to the limit in this equation shows (1.3.8).

The bound (1.3.10) also shows the left side of (1.3.9). And the right side in (1.3.9) follows from

$$\|(I - A)^{-1} - I + A\| \ \leq \ \sum_{k=2}^{\infty} \|A\|^k \ .$$

□

1.3.20 Corollary. *Let $A, B \in \mathscr{B}(X)$, assume A^{-1} exists and belongs to $\mathscr{B}(X)$ (i.e., is bounded), and assume $\|B\| < 1/\|A\|$. Then $(A + B)^{-1}$ exists and*

$$(A + B)^{-1} \ = \ A^{-1} \sum_{k=0}^{\infty} (-BA^{-1})^k \ = \ \Big(\sum_{k=0}^{\infty} (-A^{-1}B)^k \Big) A^{-1} \ .$$

Proof. Write

$$(A + B)^{-1} \ = \ [A(I + A^{-1}B)]^{-1} \ = \ [(I + BA^{-1})A]^{-1}$$

and use the fact that for any two invertible operators L and M, $(LM)^{-1} = M^{-1}\, L^{-1}$. □

1.3.21 Exercise. Let the linear mapping $\lambda : \mathbb{R}^n \to \mathbb{R}$ be given by the row vector $\lambda = (\lambda_1, ..., \lambda_n)$. Show that the operator norm $\|\lambda\| = (\sum_1^n \lambda_j^2)^{1/2}$. Work the same problem when $\lambda : \mathbb{C}^n \to \mathbb{C}$.

1.3.22 Exercise. Let $L : C([-1, 1]) \to \mathbb{C}$ be given by $Lf = f(0)$ as in Example 1.3.13, let $|\cdot|_0$ be the sup-norm, and let $r > 0$ be given. (i) Show that L is continuous using the $\epsilon\delta$-definition of continuity. (ii) Show that the inverse image $L^{-1}((-r, r)) = \{f \in C([-1, 1]) \ ; \ -r < f(0) < r\}$ is open in $C([-1, 1])$. (iii) Show that $f_n(0) \to f(0)$ whenever $f_n \to f$ in $C([-1, 1])$.

1.3.23 Exercise. Prove that $\ell^2(\mathbb{N})$ is a Hilbert space.

1.3.24 Exercise. Prove that $\mathscr{B}(X, Y)$ is a vector space when X and Y are vector spaces.

1.3.25 Exercise. Write out the details for the proof of Corollary 1.3.20.

1.3.26 Exercise. Use Corollary 1.3.20 to show that the set of invertible elements in $\mathscr{B}(X)$ is an open subset of $\mathscr{B}(X)$.

1.3.27 Exercise. Prove Proposition 1.3.3.

1.3.28 Exercise. Show that any subspace of any vector space is a convex set.

1.3.29 Exercise. Let $A \subset \mathbb{R}^n$ have the following two proeprties: (a) A is dense in $\mathbb{R}^n$, and (b) A is convex, i.e., the point $tx + (1-t)y$ lies in A whenever $x \in A$, $y \in A$, and $0 \leq t \leq 1$. Show that $A = \mathbb{R}^n$. Does this result hold if $\mathbb{R}^n$ is replaced by $\mathbb{C}^n$?

1.3.30 Exercise. Show that the set $\mathscr{P}$ of polynomials, restricted to $[0,1]$, of any degree is a dense convex subset of $C[0,1]$ (with the sup norm), but that $\mathscr{P} \neq C[0,1]$.

1.3.31 Exercise. Show that the set $\ell_0^2(\mathbb{N})$ of all finite sequences $x = (x_1, x_2, ...)$ in $\ell^2(\mathbb{N})$, having only a finite number of non-zero components, is a dense convex subset of $\ell^2(\mathbb{N})$, but that $\ell_0^2(\mathbb{N}) \neq \ell^2(\mathbb{N})$.

1.4 Examples of Linear Transformations

Differential operators

1.4.1 Example. Let $m \in \mathbb{N}$ and

$$L \;=\; \sum_{|\alpha| \leq m} a_\alpha \, \partial^\alpha \;=\; \sum_{\alpha_1 + \cdots + \alpha_d \leq m} a_\alpha \, (\frac{\partial}{\partial x_1})^{\alpha_1} \cdots (\frac{\partial}{\partial x_d})^{\alpha_d}$$

be a linear partial differential operator with constant coefficients a_α. The operator

$$L : C^m(K) \to C(K)$$

is bounded when K is any compact subset of $\mathbb{R}^d$. To show this it suffices to show that $\partial^\alpha : C^m(K) \to C(K)$ is bounded for each $|\alpha| \leq m$. This fact is established by writing

$$\sup_{x \in K} |\partial^\alpha f(x)| \;\leq\; \sum_{|\beta| \leq m} \sup_{x \in K} |\partial^\beta f(x)| \; .$$

1.4.2 Example. Let $m \in \mathbb{N}$ and

$$L = \sum_{|\alpha| \leq m} a_\alpha(x) \, \partial^\alpha \tag{1.4.1}$$

be a linear partial differential operator with variable coefficient functions $a_\alpha(x)$. Assume $a_\alpha \in C(K)$ for every α, for some compact subset $K \subset \mathbb{R}^d$. Then the operator

$$L : C^m(K) \to C(K)$$

is bounded. For we know from the preceeding example that $\partial^\alpha : C^m(K) \to C(K)$ is bounded for each $|\alpha| \leq m$. And we can show that the multiplication operators

$$a_\alpha : C(K) \to C(K) \; ,$$

given by $f(x) \mapsto a_\alpha(x) \, f(x)$, are bounded. Indeed, if $M > 0$ is so large that $\sup_{x \in K} |a_\alpha(x)| \leq M$ for all α, we have $a_\alpha(\cdot) \, f(\cdot) \in C(K)$ from the theory of continuous functions and

$$\sup_{x \in K} |a_\alpha(x) \, f(x)| \;\leq\; M \, \sup_{x \in K} |f(x)| \; .$$

Finally, for each α, $a_\alpha(x) \, \partial^\alpha \, f(x)$ is the composition of two bounded operators applied to f, and is thus bounded.

1.4.3 Example. Let L be the differential operator of Example 1.4.2. Assume that $a_\alpha(\cdot) \in L^\infty(\Omega)$ for some open $\Omega \subset \mathbb{R}^d$. (Here, $L^\infty(\Omega)$ is the vector space of bounded functions on Ω.) We claim that

$$L : H^m(\Omega) \to H^0(\Omega) = L^2(\Omega)$$

is bounded. It suffices to show that, for every $|\alpha| \le m$, both operators

$$\partial^\alpha : H^m(\Omega) \to L^2(\Omega) \ ,$$

and

$$a_\alpha : L^2(\Omega) \to L^2(\Omega)$$

are bounded. The truth of these statements follows from the bounds

$$\int_\Omega |\partial^\alpha f(x)|^2 \, dx \ \le \ \sum_{|\beta| \le m} \int_\Omega |\partial^\beta f(x)|^2 \, dx$$

and

$$\int_\Omega |a_\alpha(x) \, f(x)|^2 \, dx \ \le \ M^2 \int_\Omega |f(x)|^2 \, dx \, ,$$

where $M > 0$ is chosen so large that $\sup_{x \in \Omega} |a_\alpha(x)| \le M$ for all α.

Integral operators

1.4.4 Example (L^2 kernels)**.** Let $\Omega \subset \mathbb{R}^d$ be open and $k \in L^2(\Omega \times \Omega)$. Define the linear transformation $K : L^2(\Omega) \to L^2(\Omega)$ by

$$Kf(x) \ = \ \int_\Omega k(x,y) f(y) \, dy \ . \tag{1.4.2}$$

K is bounded because

$$\Big| \int_\Omega k(x,y) f(y) \, dy \Big| \ \le \ \Big(\int_\Omega |k(x,y)|^2 \, dy \ \int_\Omega |f(y)|^2 \, dy \Big)^{1/2}$$

by Schwarz' inequality, and therefore

$$\int_\Omega |Kf(x)|^2 \, dx \ \le \ \int_\Omega \int_\Omega |k(x,y)|^2 \, dy \, dx \ \int_\Omega |f(y)|^2 \, dy \ .$$

This shows that $\|K\| \le \big(\int_\Omega \int_\Omega |k(x,y)|^2 \, dy \, dx \big)^{1/2}$.

1.4.5 Example (Fourier transform)**.** Let $C_b(\mathbb{R}^d)$ be the vector space of continuous functions on $\mathbb{R}^d$ which are bounded on all of $\mathbb{R}^d$. And put on this space the norm $|f|_0 = \sup_{x \in \mathbb{R}^d} |f(x)|$ (uniform convergence on all of $\mathbb{R}^d$). Let $T : L^1(\mathbb{R}^d) \to C_b(\mathbb{R}^d)$ be defined by

$$Tf(x) \ = \ \int_{\mathbb{R}^d} e^{-ix \cdot y} f(y) \, dy \ .$$

The kernel $e^{-ix \cdot y}$ in this example is not in $L^2(\mathbb{R}^d)$. But T is bounded because

$$|Tf(x)| \ \le \ \int_{\mathbb{R}^d} |e^{-ix \cdot y}| \, |f(y)| \, dy \ \le \ \int_{\mathbb{R}^d} |f(y)| \, dy$$

for every $x \in \mathbb{R}^d$. Thus $\sup_x |Tf(x)| \le \int_{\mathbb{R}^d} |f(y)| \, dy$, and $\|T\| \le 1$.

Linear functionals All of the following linear transformations are of the form $L : X \to \mathbb{K}$ for some Banach space X.

1.4.6 Example (integration)**.** Let $\Omega \subset \mathbb{R}^d$ be open and $L^1(\Omega)$ be the vector space of integrable functions on Ω, that is, the set of functions f such that $\int_\Omega |f(x)|\,dx < \infty$. And let $L : L^1(\Omega) \to \mathbb{K}$ be defined by $Lf = \int_\Omega f(x)\,dx$. Then L is bounded since $|\int_\Omega f(x)\,dx| \le \int_\Omega |f(x)|\,dx$.

1.4.7 Example (evaluation)**.** Let $X = H^1((-1,1))$. We will show that the evaluation linear functional $u \mapsto u(0)$ is continuous on X. To say it another way, the Dirac delta function δ is a bounded linear functional on X. If $u \in X$ the value of $\delta(u)$ will not be changed if we multiply u by a $\phi \in C_0^\infty((-1,1))$ which satisfies $0 \le \phi \le 1$ on $(-1,1)$ and $\phi \equiv 1$ on $(-0.5, 0.5)$ for instance. And because X is the completion of $C^\infty(-1,1)$ with respect to $\|u\|_1 = \int_{-1}^1 |u|^2 + |u'|^2\,dx$ it suffices to show that δ is a bounded linear functional on $C_0^\infty(-1,1)$ with respect to the norm $\|\cdot\|_1$. For if u is any element of X and u_n is a sequence in $C_0^\infty(-1,1)$, we may define $u(0) = \lim_n u_n(0)$ if this linear functional is continuous on a dense subset. So, for $u \in C_0^\infty(-1,1)$ we observe that $u(x) = \int_{-1}^x u'(s)\,ds$. Then

$$|u(x)| \le \int_{-1}^x |u'(s)|\,ds \le \left(\int_{-1}^x 1\,ds \int_{-1}^x |u'(s)|^2\,ds\right)^{1/2} \le \sqrt{2}\left(\int_{-1}^x |u'(s)|^2\,ds\right)^{1/2}$$

by applying the Schwartz inequality to $\int_{-1}^x 1\,|u'|$. Since $(\int_0^x |u'(s)|^2\,ds)^{1/2} \le \|u\|_1$ we have $|u(x)| \le \sqrt{2}\|u\|_1$ for every $x \in (-1,1)$. We have shown in particular that the operator norm $\|\delta\| \le \sqrt{2}$.

1.4.8 Example. Let H be any Hilbert space with inner product $\langle\cdot,\cdot\rangle$. If we fix any $y \in H$ the mapping $x \mapsto \langle y, x\rangle$ is a linear functional on H. The Cauchy-Schwarz inequality shows that this functional is bounded: $|\langle y, x\rangle| \le M\|x\|$ where $M = \|y\|$.

For instance, if $\Omega \subset \mathbb{R}^d$ is open and $g \in L^2(\Omega)$ is fixed, the functional $f \mapsto \int_\Omega f(x)g(x)\,dx$ is a bounded linear functional on $L^2(\Omega)$.

Fourier series

1.4.9 Example (Fourier series)**.** For any $f \in L^2(0, 2\pi)$ the Fourier series expansion of f can be written as

$$f(x) \;=\; \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos nx \;+\; b_n \sin nx)$$

where the Fourier coefficients are

$$a_n \;=\; \frac{1}{\pi}\int_0^{2\pi} f(x)\,\cos(nx) \text{ and } b_n \;=\; \frac{1}{\pi}\int_0^{2\pi} f(x)\,\sin(nx)$$

for $n = 0, 1, 2, \ldots$. The mapping $L : L^2(0, 2\pi) \to \ell^2$ given by $f \mapsto (a_0, a_1, b_1, a_2, b_2, \ldots)$ is a linear transformation. By Parseval's theorem

$$\int_0^{2\pi} |f(x)|^2\,dx \;=\; |a_0|^2 + \sum_{1}^{\infty}(|a_n|^2 + |b_n|^2)\;.$$

Taking the square root of both sides of this equation shows that $\|L\| = 1$; in fact it is well known from the theory of Fourier series that L is an isometry of $L^2(0, 2\pi)$ onto ℓ^2.

1.4.10 Exercise. Fix $k \in \mathbb{N}$ and let L be the differential operator in Example 1.4.1. Show that $L : C^{m+k}(K) \to C^k(K)$ is a bounded linear transformation.

1.4.11 Exercise. Let L be the differential operator of Example 1.4.2. Find an upper bound for the operator norm $\|L\|$.

1.4.12 Exercise. Let L be the differential operator of Example 1.4.3. Find an upper bound for the operator norm $\|L\|$.

1.4.13 Exercise. Let L be given by (1.4.1), except that each $a_\alpha(x)$ is an $m \times n$ matrix valued function whose components are in $L^\infty(\Omega)$. Show that

$$L : H^m(\Omega; \mathbb{C}^n) \to H^0(\Omega; \mathbb{C}^m) = L^2(\Omega; \mathbb{C}^m)$$

is a bounded linear transformation.

Assume now that Ω is bounded and that the components of $a_\alpha(x)$ belong to $C(\overline{\Omega})$. Show that

$$L : C^m(\overline{\Omega}; \mathbb{C}^n) \to C(\overline{\Omega}; \mathbb{C}^m)$$

is a bounded linear transformation.

1.4.14 Exercise. Let $L^1(\mathbb{R})$ be the vector space of integrable functions on $\mathbb{R}$. ($f \in L^1(\mathbb{R})$ if $\int_{-\infty}^{\infty} |f(x)|\, dx < \infty$.) Fix $g \in L^1(\mathbb{R})$, and show that $Lf = \int_{-\infty}^{\infty} f(y)g(x-y)\, dy$ defines a bounded linear transformation $L : L^1(\mathbb{R}) \to L^1(\mathbb{R})$. Show also that $\|L\| \le \int_{-\infty}^{\infty} |g(x)|\, dx$.

1.5 Projections in Hilbert Space

It is the notion of orthogonality that distinguishes inner product spaces from other normed spaces. The linear transformations discussed in this section exploit this property.

Closed subspaces Recall that a subset of a vector space, which is itself a vector space, is called a *subspace*. If X and Y are vector spaces over the same field, $\mathbb{R}$ or $\mathbb{C}$, and $L : X \to Y$ is a linear function, then the null space of L, $N(L) = \{x \in X \ ; \ Lx = 0\}$, is a subspace of X, and the range space of L, $R(L) = \{y \in Y \ ; \ y = Lx \text{ for some } x \in X\}$, is a subspace of Y.[5]

If the vector space has a norm, and if the subspace is closed, we speak of a *closed subspace*. Closed subspaces are important because the projection of a vector onto a subspace may not exist if the subspace is not cloased. A good analogy is the 'projection' of the vector $(0,1)$ in $\mathbb{R}^2$ onto the diagonal $\{(x,y) \ ; \ y = x \text{ and } x, y \in \mathbb{Q}\}$, with only rational coordinates. The 'projection' is the point $(1/\sqrt{2}, 1/\sqrt{2})$ but this point is not in the subset which forms a vector space over $\mathbb{Q}$.

In a finite dimensional vector space every subspace is closed, but this is false in general. For instance the vector space of polynomials $\mathscr{P}[0,1]$ (of any degree) on $[0,1]$ is a subspace of $C([0,1])$. By the Weierstrass approximation theorem the closure of this subspace is all of $C([0,1])$, so $\mathscr{P}[0,1]$ is not closed in $C([0,1])$.

If X and Y are normed vector spaces and L is continuous, then $N(L)$ is closed in X because it is the inverse image of a closed set, $\{0\}$, in Y. (But $R(L)$ need not be closed in Y even when L is continuous).

By Proposition 1.1.19, every finite dimensional subspace of any normed vector space is closed.

Normal equations

1.5.1 Lemma. *Let H be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, let the finite set of vectors $x_1, x_2, ..., x_m$ be linearly independent in H and $X \subset H$ their linear span, and let y be any vector in H. Then there exists a unique $\tilde{y} \in X$ such that $(y - \tilde{y}) \perp x_i$ for every $i = 1, 2, ..., m$. If we set $\tilde{y} = x_1\tilde{\alpha}_1 + \cdots x_m\tilde{\alpha}_m$, the $\tilde{\alpha}_j$'s are solutions of the linear system*

$$\begin{pmatrix} \langle x_1, x_1\rangle & \langle x_1, x_2\rangle & \cdots & \langle x_1, x_m\rangle \\ \langle x_2, x_1\rangle & \langle x_2, x_2\rangle & \cdots & \langle x_2, x_m\rangle \\ \vdots & \vdots & & \vdots \\ \langle x_m, x_1\rangle & \langle x_m, x_2\rangle & \cdots & \langle x_m, x_m\rangle \end{pmatrix} \begin{pmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \\ \vdots \\ \tilde{\alpha}_m \end{pmatrix} = \begin{pmatrix} \langle x_1, y\rangle \\ \langle x_2, y\rangle \\ \vdots \\ \langle x_m, y\rangle \end{pmatrix}. \tag{1.5.1}$$

Recall that the $m \times m$ matrix $G = [\langle x_i, x_j\rangle]$ is called the Gram matrix of the x_j's.

The system (1.5.1) is called the *normal equations*.

[5] Note that this is an example of the fact that both the kernel and image of any group homomorphism are subgroups.

Proof. The condition $x_i \perp (y - \tilde{y})$ means precisely that

$$\langle x_i, y - x_1\tilde{\alpha}_1 - x_2\tilde{\alpha}_2 - \cdots - x_m\tilde{\alpha}_m \rangle \;=\; 0\,;$$

this is just the i-th row of the system (1.5.1). And, any $\tilde{y}$ satisfying such a perpendicular condition must also satisfy these equations. Since the x_j's are linearly independent the Gram matrix is non-singular by Proposition 1.1.15. So this system of equations has a unique solution. □

The proof of the next theorem uses the Pythagorean theorem (Exercise 1.5.19): if x and y are vectors in H such that $\langle x, y\rangle = 0$, then $\|x + y\|^2 = \|x\|^2 + \|y\|^2$.

1.5.2 Theorem. *Let H be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, let the finite set of vectors $x_1, x_2, ..., x_m$ be linearly independent in H and $X \subset H$ their linear span, and let y be any vector in H. Define the squared distance function $f : X \to \mathbb{R}$ by*

$$f(x) \;=\; \|y - x\|^2\,.$$

Then there exists a unique $\hat{y} \in X$ which minimizes $f(x)$ over all $x \in X$. The minimizer $\hat{y} = x_1\hat{\alpha}_1 + \cdots + x_m\hat{\alpha}_m$ is given by $\hat{\alpha}_j = \tilde{\alpha}_j$, $j = 1, ..., m$, the solutions of the normal equations (1.5.1).

Proof. Choose any $z \in X$ and set $R = \|y - z\|$. Clearly $f(x)$ has an infimum ≥ 0, so $0 \leq \inf\{f(x)\ ;\ x \in X\} \leq R^2$. Now X is a finite dimensional vector space, with topology the same as K^m by Proposition 1.1.19. So $B = \{x \in X\ ;\ \|y - x\| \leq R\}$ is closed and bounded, hence compact. And f is continuous on B since $\|y - x\|^2 \leq (\|y\| + \|x\|)\|y - x\| \leq 2(\|y\| + R)\|y - x\|$ when $x \in B$. So f assumes its minimum on B, at some point we call $\hat{y}$.

We must now show that $\hat{y} = \tilde{y}$ where $\tilde{y}$ is the vector in X that makes $y - \tilde{y}$ orthogonal to the subspace X (Lemma 1.5.1). Suppose $\hat{y} \neq \tilde{y}$. Then $(\hat{y} - \tilde{y}) \perp (y - \tilde{y})$ since $\hat{y} - \tilde{y} \in X$. So the Pythagorean theorem implies that

$$\|y - \hat{y}\|^2 \;=\; \|y - \tilde{y}\|^2 + \|\tilde{y} - \hat{y}\|^2 \;>\; \|y - \tilde{y}\|^2\,.$$

But this contradicts the definition of $\hat{y}$ as the minimizer of $f(x)$. So $\hat{y} = \tilde{y}$. □

Orthogonal subspaces

1.5.3 Definition. If X is an inner product space and $x, y \in X$ we say x and y are *orthogonal* if $\langle x, y\rangle = 0$, and write $x \perp y$. If U is a subset of X we define the *orthogonal subspace*

$$U^\perp = \{x \in X\ ;\ \langle x, y\rangle = 0 \quad \text{for all} \quad y \in U\}.$$

If U and V are subsets of X, we write $U \perp V$ to mean that $\langle x, y\rangle = 0$ whenever $x \in U$ and $y \in V$. The notation $x \perp V$ or $V \perp x$ means that $\langle x, y\rangle = 0$ whenever $y \in V$.

The zero vector is the only element orthogonal to itself. It is easy to see that $X^\perp = \{0\}$ and that $\{0\}^\perp = X$.

1.5.4 Lemma. *If U is any subset of a Hilbert space H, $U^\perp$ is a closed subspace of H.*

Proof. In Exercise 1.5.18 we ask the reader to show that $U^\perp$ is a subspace of H. We will show that $U^\perp$ is complete.

Let x_n be a Cauchy sequence in $U^\perp$ and $x_n \to x_0$ in H. We must show that $x_0 \in U^\perp$. For any $y \in U$ we have $\langle y, x_n\rangle = 0$, so by Schwarz inequality

$$|\langle y, x_0\rangle| \;=\; |\langle y, x_0\rangle - \langle y, x_n\rangle| \;\leq\; \|y\|\,\|x_0 - x_n\|\,.$$

Since the right hand side converges to zero as $n \to \infty$, $\langle y, x_0\rangle = 0$. □

1.5.5 Theorem. *If V is a closed subspace of a Hilbert space H then $H = V \oplus V^\perp$. More precisely the statement means that every $y \in H$ has a unique decomposition $y = x + z$ where $x \in V$ and $z \in V^\perp$. In this case $\|y\|^2 = \|x\|^2 + \|z\|^2$.*

Proof. The proof of the general case can be found in Rudin, *Real and Complex Analysis*, Theorem 4.11. Here we prove the theorem when V is finite dimensional.

Let $\tilde{y} \in V$ be the unique vector such that $\tilde{y} - y \perp V$ (Lemma 1.5.1). Then the vectors $x = \tilde{y}$ and $z = y - \tilde{y}$ satisfy the claims of the theorem. □

Projections

1.5.6 Definition. Let H, V, y, x, and z be as in Theorem . We call the functions $P : H \to H$ and $Q : H \to H$, given by

$$Py \;=\; x \qquad \text{and} \qquad Qy \;=\; z\,,$$

the *projections* of H onto V and $V^\perp$, respectively.

1.5.7 Theorem. *The projections P and Q in the preceeding definition have the following properties.*

a) $\mathscr{R}(P) \;=\; V \quad$ *and* $\quad \mathscr{N}(P) = V^\perp$

b) $P^2 \;=\; P$

c) $Q \;=\; I - P$

d) $\mathscr{R}(Q) \;=\; V^\perp \quad$ *and* $\quad \mathscr{N}(Q) = V$

e) $Q^2 \;=\; Q$

f) If $\dim(V) < \infty$ *and* $V = \operatorname{span}\{x_1, ..., x_n\}$*, then*

$$Py \;=\; \sum_1^n x_i \tilde{\alpha}_i$$

where $\tilde{\alpha}_1$, ..., $\tilde{\alpha}_n$ are the solutions of the normal equations (1.5.1).

Proof. The proof is a matter of routine checking, or application of the previous results, and we leave it as an exercose for the reader. □

1.5.8 Example. Let $H = L^2(0,1)$ and V be the span of $1, x, x^2, ..., x^{m-1}$. Then the ij-th entry of the Gram matrix is $\int_0^1 x^{i+j-2}\,dx$, and if $y = y(x)$ is any square integrable function on $(0,1)$ the vector b has components $\int_0^1 x^{j-1}\,y(x)\,dx$. The polynomial of degree $\le m-1$ which is closest to the function $y(x)$, in the L^2 sense on $(0,1)$, is $\sum_{j=0}^{m-1} x^{j-1}\alpha_j$ where the $\mathbb{K}^m$ valued vector $a = (\alpha_1, ..., \alpha_m)' = G^{-1}b$.

1.5.9 Example. Let H be the complex Hilbert space $L^2(0,1)$ and V be the $m = 2p+1$ dimensional subspace spanned by $\{e^{2\pi ikx} \;;\; -p \le k \le p\}$. This set is already orthonormal with respect to the inner product $\langle u, v\rangle = \int_0^1 \bar{u}(x)\,v(x)\,dx$. So the Gram matrix G is $I_{n\times n}$. For any $y \in H$ we have $\hat{y}(x) = \sum_{k=-p}^{p} e^{2\pi ikx} b_k$ where b is the vector of Fourier coefficients $b_k = \int_0^1 e^{-2\pi ikx}\,y(x)\,dx$.

Complete orthonormal sets All Hilbert spaces that arise in applications are separable, and the assumption of separability makes certain computations easier.

1.5.10 Definition. A metric space (X, d) is *separable* if it has a countable dense subset. That is, if there is a set $Z = \{z_1, z_2, z_3, ...\} \subset X$ such that for every $\epsilon > 0$ and every $x \in X$, there is a $z_j \in Z$ for which $d(x, z_j) < \epsilon$.

1.5.11 Definition. A set $e_1, e_2, e_3, ...$ in a separable Hilbert space H is an *orthonormal basis* or a *complete orthonormal set* (c.o.n.s.) if the following two properites hold:

a) $\langle e_i, e_j\rangle \;=\; \delta_{ij}$ for all i and j in $\mathbb{N}$, and

b) every $x \in H$ has an expansion of the form

$$x \;=\; \sum_{j=1}^{\infty} \alpha_j\, e_j \tag{1.5.2}$$

with coefficients $\alpha_j \in \mathbb{K}$, where the series converges to x in H. This means that $\|x - s_n\| \to 0$ as $n \to \infty$ where $s_n = \sum_{j=1}^n \alpha_j\, e_j$ is the sequence of partial sums.

Since H is complete (1.5.2) is convergent in H if and only if the sequence s_n is Cauchy.

1.5.12 Lemma. *The series (1.5.2) is convergent in H if and only if $\sum_{j=1}^{\infty} |\alpha_j|^2 < \infty$.*

Proof. Let $m, n \in \mathbb{N}$ with $m < n$. Set $x_m = \sum_1^m \alpha_j\, e_j$ and similarly for x_n. Since $\langle e_i, e_j\rangle = \delta_{ij}$ we have

$$\|x_n - x_m\|^2 \;=\; \langle x_n - x_m, x_n - x_m\rangle \;=\; \sum_{m+1}^{n} |\alpha|^2 \;.$$

Thus the sequence $\|x_n\|$ is Cauchy in H if and only if the partial sums of the series $\sum_1^\infty |\alpha_j|^2$ form a Cauchy sequence in $\mathbb{R}$. □

1.5.13 Proposition. *Every separable Hilbert space H has a countable, complete orthonormal set.*

Proof. Let $\langle\cdot,\cdot\rangle$ denote any inner product on H. By the separability assumption there is a countable dense set $v_1, v_2, ...$ in H. Assume, without loss of generality, that $v_1 \neq 0$ and set $e_1 = v_1/\|v_1\|$.

Next, if v_2 is either zero or in the linear span of v_1 we discard it. If not, set $u_2 = v_2 - e_1\,\langle e_1, v_2\rangle$, and then $e_2 = u_2/\|u_2\|$. The vector u_1 is the projection of v_1 onto $e_1^\perp$, and the construction of e_2 is the first step in the Gram-Schmidt orthogonalization process.

Now we proceed inductively to construct the sequence e_n. Assume the orthonormal set $e_1, ..., e_{n-1}$ has been constructed from $v_1, .., v_{m-1}$. (Notice that $m \geq n$ and v_j is in the linear span of $e_1, ..., e_{n-1}$ when $j \leq m-1$.) If v_m is either zero or in the linear span of $e_1, ..., e_{n-1}$, discard it and consider v_{m+1}. (If we renumber the v's after discarding one, we may assume $m = n$.) If v_m is not zero or in the span, set

$$u_n \;=\; v_m - e_1\,\langle e_1, v_m\rangle - e_2\,\langle e_2, v_m\rangle - \cdots - e_{n-1}\,\langle e_{n-1}, v_m\rangle\,.$$

This is the projection of v_m onto $\{e_1, ..., e_{n-1}\}^\perp$. Then set $e_n = u_n/\|u_n\|$.

In this way we construct an infinite (or finite if H is finite dimensional) sequence e_n of vectors in H which satisfy $\langle e_i, e_j\rangle = \delta_{ij}$ when $i, j \in \mathbb{N}$.

The last step in the proof is to show that every $u \in H$ is the limit of finite linear combinations of the e_j's. To see this we first observe that the scalars α_j which minimize $\|u - \sum_1^n \alpha_j\, e_j\|$ for a fixed n are $\alpha_j = \langle e_j, u\rangle$. This follows from the normal equations since the Gram matrix is the identity.

Now observe that the construction of the e_j's shows that every element v_n of the original dense set in H can be written as a linear combination of the set $e_1, ..., e_n$. So let $v_n = \sum_1^n \gamma_j\, e_j$ (the γ_j's actually depend on n as well but we suppress this in the notation).

For any $\epsilon > 0$ one can choose an $n_0 \in \mathbb{N}$, sufficiently large, so that $n \geq n_0$ implies $\|u - v_n\| < \epsilon$. For any such n, $\|u - \sum_1^n \langle e_j, u\rangle\, e_j\| \leq \|u - \sum_1^n \gamma_j\, e_j\| = \|u - v_n\| < \epsilon$. Thus, $\|u - \sum_1^n \langle e_j, u\rangle\, e_j\| \to 0$ as $n \to \infty$. □

1.5.14 Lemma (Parseval's identity)**.** *If H is a separable Hilbert space with c.o.n.s. e_n, and $x = \sum_1^\infty a_n e_n$ and $y = \sum_1^\infty b_n e_n$ are any two vectors in H, then*

$$\langle x, y\rangle \;=\; \sum_{n=1}^{\infty} a_n \bar{b}_n\,.$$

Proof. We only exhibit this result formally, that is, by freely exchanging limits. We have

$$\langle \sum_{m=1}^{\infty} a_m e_m, \sum_{n=1}^{\infty} b_n e_n\rangle \;=\; \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} a_m \bar{b}_n \langle e_m, e_n\rangle \;=\; \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} a_m \bar{b}_n \delta_{mn} \;=\; \sum_{n=1}^{\infty} a_n \bar{b}_n\,.$$

□

1.5.15 Definition. Two Hilbert spaces H_1 and H_2 are *isomorphic* if there is a one-to-one, onto, linear function (called an 'isomorphism') $\phi : H_1 \to H_2$ which satisfies $\langle x, y\rangle_1 = \langle \phi(x), \phi(y)\rangle_2$ for all x and y in H_1.

1.5.16 Corollary. *Every separable Hilbert space H is isomorphic to $\ell^2(\mathbb{J})$ where $\mathbb{J}$ is any countable index set.*[6]

1.5.17 Exercise. Let X and Y be vector spaces over the same field, $\mathbb{R}$ or $\mathbb{C}$, and $L : X \to Y$ be a linear transformation. Show that the null space of L, $N(L) = \{x \in X\ ;\ Lx = 0\}$, is a subspace of X. Show that the range space of L, $R(L) = \{y \in Y\ ;\ y = Lx \text{ for some } x \in X\}$, is a subspace of Y.

[6] Typical examples of $\mathbb{J}$ include $\mathbb{N}, \mathbb{Z}, \mathbb{N}^k$, and $\mathbb{Z}^k$.

1.5.18 Exercise. Assume that X is an inner product space and that $U \subset X$ is any subset. Show that $U^\perp$ is a subspace of X.

1.5.19 Exercise (Pythagorean theorem)**.** If x, y are elements of an inner product space X and if $x \perp y$, expand $\langle x \pm y, x \pm y\rangle$ to show that

$$\|x \pm y\|^2 = \|x\|^2 + \|y\|^2.$$

Assume now that $x_1, ..., x_n$ are mutually orthogonal in X, i.e., that $\langle x_i, x_j\rangle = 0$ if $i \neq j$. Show that

$$\|x_1 + \cdots + x_n\|^2 \;=\; \|x_1\|^2 + \cdots + \|x_n\|^2 .$$

1.5.20 Exercise. Let U and V be two closed subspaces of a Hilbert space, and assume $U \perp V$. Show that $U + V$ is closed.

Solution: Let y_n be Cauchy in $U + V$, and $y_n = x_n + z_n$ with $x_n \in U$ and $z_n \in V$. Since $U \perp V$ the Pythagorean theorem implies $\|y_n - y_m\|^2 = \|x_n - x_m\|^2 + \|z_n - z_m\|^2$ so that x_n is Cauchy in U and z_n is Cauchy in V. If $x_n \to x$ and $z_n \to z$ then, again from the Pythagorean theorem, $y_n \to y = x + z \in U + V$.

1.5.21 Exercise. Let H be a Hilbert space and $P : H \to H$ have the property $P^2 = P$. Show that $Q = I - P$ also satisfies $Q^2 = Q$.

1.5.22 Exercise. Let H be a Hilbert space and $z \in H$ any non-zero vector. Show that the projection of $x \in H$ onto the span of z is

$$\tfrac{z}{\|z\|}\langle \tfrac{z}{\|z\|}, x\rangle .$$

Let z_1 and z_2 be non-zero orthogonal vectors in H. Show that the projection of $x \in H$ onto the span of z_1 and z_2 is

$$\tfrac{z_1}{\|z_1\|}\langle \tfrac{z_1}{\|z_1\|}, x\rangle + \tfrac{z_2}{\|z_2\|}\langle \tfrac{z_2}{\|z_2\|}, x\rangle .$$

1.5.23 Exercise. Prove Theorem 1.5.7.

1.6 Duality and the Test Function Principle

When the range space of the linear functions discussed in the last section is the scalar field some additional useful results can be obtained. These functions play an extremely important role in many applications. If $0 \leq x \leq 1$ and $f \in C([0, 1])$ the mapping

$$\lambda_x(f) = f(x)$$

which evaluates the continuous function f at x is linear *and* bounded (i.e., continuous).[7] Linear transformations of this form are very important to us. Indeed we usually think of functions (at least contiuous ones) as defined by the values they take on their domain! This simple example illustrates how important linear functionals are to us, perhaps even without our being aware of it.

A second example is the coordinate functionals on $\mathbb{K}^n$. Consider $\mathbb{R}^n$, for definiteness, and let's take its elements to be column n-vectors. All linear functionals, linear transformations from $\mathbb{R}^n$ to $\mathbb{R}$, can be given by the set of all $1 \times n$ matrices with real components, that is, by the set of all *row* n-vectors. These linear functionals (e.g., those of unit length) pick off the components of the column vectors. In this way one can view the linear functionals on $\mathbb{R}^n$ as being exactly the set of linear coordinates by which the vector space $\mathbb{R}^n$ may be parameterized.

We begin our discussion with a simple, but very important, proposition.

1.6.1 Proposition. *If x and y are two elements of an inner product space X, then $x = y$ if and only if $\langle z, x\rangle = \langle z, y\rangle$ for all $z \in X$.*

Proof. The 'only if' part is obvious from the definition of the inner product as a function.

To prove the 'if' part, let x and y be given and assume $\langle z, x\rangle = \langle z, y\rangle$ for all $z \in X$. This implies that $\langle z, x\rangle - \langle z, y\rangle = \langle z, x - y\rangle = 0$ for all z, and setting $z = x - y$ gives $\|x - y\|^2 = 0$ or $x = y$. □

[7] The same linear mapping on $L^2([0, 1])$ instead of $C([0, 1])$ is no longer continuous.

Here is a useful extension of the proposition.

1.6.2 Corollary. *Let x and y be two elements of an inner product space X. Then $x = y$ if and only if $\langle z, x\rangle = \langle z, y\rangle$ for all z in a dense subset of X.*

Proof. The 'only if' is obvious. To show the 'if' part, we need to use the fact that the inner product is a continuous function, that is, we need to know that $\lim_n \langle z_n, x\rangle = \langle z, x\rangle$ if $z_n \to z$ as $n \to \infty$. One sees this from the Schwarz inequality: $|\langle z_n - z, x\rangle| \le \|z_n - z\|\,\|x\|$.

Now let Z be the dense subset of X, z be any point of X, and z_n a sequence in Z converging to z. We must show that $\langle z, x\rangle = \langle z, y\rangle$. But we have by hypothesis $\langle z_n, x\rangle = \langle z_n, y\rangle$ for all n. Passing to the limit on both sides of these equalities gives $\langle z, x\rangle = \langle z, y\rangle$. □

1.6.3 Example. Let ℓ^2 be the Hilbert space of square summable sequences, $x = (x_1, x_2, ...)$ such that $\sum_1^\infty |x_j|^2 < \infty$. A sequence x is clearly determined by the numerical values $\langle y, x\rangle = \sum_1^\infty \overline{y}_j x_j$ for every $y \in \ell^2$. For each 'natural basis' vector $e_n = (0, ..., 0, 1, 0, ...)$ with a 1 in the n-th place is in ℓ^2, and $\langle e_n, x\rangle = x_n$ gives the n-th component of x.

1.6.4 Example. By Corollary 1.6.2, two functions f and g in $L^2(\Omega)$ are equal as L^2 functions if and only if

$$\int_\Omega f(x)\phi(x)\,dx \;=\; \int_\Omega g(x)\phi(x)\,dx \qquad \text{for all} \quad \phi \in C_0^\infty(\Omega)\,.$$

Here $\Omega \subset \mathbb{R}^d$ is any open subset and the statement holds because $L^2(\Omega)$ can be defined as the completion of $C_0^\infty(\Omega)$ in the L^2 norm.

Note that equality in this context does *not* mean $f(x) = g(x)$ for every $x \in \Omega$, but that equality hold for all x in Ω except possibly for a set of measure zero.

In the process of modern analysis, the functions $\phi \in C_0^\infty(\Omega)$ in the previous example are usually called *test functions* because they are used to test equality. They play the role of the independent variable x in an expression $f(x)$. Here is another example where these functions are used.

1.6.5 Example. Since the test functions $C_0^\infty(\Omega)$ are dense in $H_0^1(\Omega)$ (by definition), we have $f = g$ in $H_0^1(\Omega)$ if and only if $\langle f, \phi\rangle_1 = \langle g, \phi\rangle_1$, or

$$\int_\Omega f\,\overline{\phi} + \partial_1 f\,\partial_1\overline{\phi} + \cdots + \partial_d f\,\partial_d\overline{\phi}\;dx \;=\; \int_\Omega g\,\overline{\phi} + \partial_1 g\,\partial_1\overline{\phi} + \cdots + \partial_d g\,\partial_d\overline{\phi}\;dx\,,$$

for every test function ϕ.

1.6.6 Definition. If X is a normed vector space over $\mathbb{K}$, $\mathscr{B}(X, \mathbb{K})$ is called the *dual* of X. For convenience it is often denoted by X'.

Since $\mathbb{K}$ is complete the following is an immediate corollary of Theorem 1.3.16.

1.6.7 Proposition. *If X is a normed vector space, X' is a Banach space when given the operator norm.*

If $\lambda : X \to \mathbb{K}$, recall that the operator norm is $\|\lambda\| = \sup\{|\lambda(x)|\ ;\ \|x\|_{\mathrm{x}} = 1\}$.

In a Hilbert space H the vector space of all bounded linear functionals on H can be very precisely identified.

1.6.8 Theorem (Riesz representation)**.** *Let H be a Hilbert space with inner product $\langle\cdot,\cdot\rangle$, and $\lambda : H \to \mathbb{K}$ be a bounded linear functional on H. Then there is a unique $y \in H$, depending on λ, such that*

$$\langle x, y\rangle \;=\; \lambda(x) \qquad \textit{for every} \quad x \in H\,. \tag{1.6.1}$$

Moreover, $\|y\| = \|\lambda\|$.

The right side of the last statement is the operator norm of λ.

Abstract proof. Existence. If $\lambda(x) = 0$ for all $x \in H$ then $y = 0$ will work. If λ is not identically zero the null space of λ, call it $N = \{x \in H \ ; \ \lambda(x) = 0\}$, is not all of H. Since λ is linear N is a subspace, and since λ is continuous N (the pull back of a closed set) is closed. Proposition TBD then implies that there is a non-zero $z \in N^\perp$.

In general, given any linear functional λ on any vector space, and any two vectors x and z, the vector $\lambda(x)z - \lambda(z)x$ always lies in the null space of λ as we see by applying λ to it. In the present situation if we take $z \in N^\perp$ with $z \neq 0$ we get $\langle \lambda(x)z - \lambda(z)x, z\rangle = 0$. This shows that $\lambda(x) = \lambda(z)\langle x, z\rangle/\langle z, z\rangle$ for every $x \in H$. So (1.6.1) is true with $y = \overline{\lambda(z)}z/\|z\|^2$.

Uniqueness. If both y_1 and y_2 represent λ then for all $x \in H$ we have $\langle x, y_1\rangle = \langle x, y_2\rangle$ or $\langle x, y_1 - y_2\rangle = 0$. Setting $x = y_1 - y_2$ shows that $y_1 - y_2 = 0$. □

Constructive proof when H is separable. Since H is separable, Proposition 1.5.13 shows that there is a complete orthonormal set $\{e_j \ ; \ j \in \mathbb{N}\}$ for H. Denote the value of λ on e_j by $\beta_j = \lambda(e_j)$. We want to find $u = \sum_1^\infty \alpha_i e_i \in H$ (i.e., with $\sum_1^\infty |\alpha_i|^2 \leq \infty$ but all α_i's unknown at the moment) such that

$$\langle u, e_j\rangle \ = \ \lambda(e_j)$$

for all $j \in \mathbb{N}$. (This will imply that $\langle u, v\rangle \ = \ \lambda(v)$ for all $v \in H$ since every such v is the limit of a finite linear combination of the e_j's, and since both $\langle \ , \ \rangle$ and λ are *continuous* in their arguments.) But this means that

$$\sum_1^\infty \bar\alpha_i \, \langle e_i, e_j\rangle \ = \ \sum_1^\infty \bar\alpha_i \, \delta_{ij} \ = \ \bar\alpha_j \ = \ \lambda(e_j) \ = \ \beta_j$$

where $\bar\alpha_j$ is the complex conjugate of α_j. (We have again used the continuity of the inner product to pull the limit outside.) This then says that we must have $\alpha_j = \bar\beta_j$ for all $j \in \mathbb{N}$, and hence solves for the required u.

Finally we must show that the u above lies in H, that is, that $\sum_1^\infty |\beta_j|^2 \leq \infty$. Of course, if all β_j's are zero this is trivial, so in the following we need only be concerned with those n which are so large that $\sum_1^n |\beta_j|^2 > 0$. We are going to use the fact that λ is bounded; this means that

$$\sup_{\|v\|=1} |\lambda(v)| \ = \ \|\lambda\|$$

is a finite real number. The supremum here is taken over all vectors in H of norm 1. Consider the sequence of vectors $v_n = \sum_1^n \bar\beta_j e_j$ in H where $n \in \mathbb{N}$ is sufficiently large. We can rescale these to have length 1, so we actually will apply λ to the vectors

$$v_n = \frac{1}{(\sum_1^n |\beta_j|^2)^{1/2}} \sum_{j=1}^n \bar\beta_j e_j.$$

We have

$$\begin{aligned} \lambda(v_n) \ &= \ (\sum_{j=1}^n |\beta_j|^2)^{-1/2} \sum_{j=1}^n \bar\beta_j \lambda(e_n) \\ &= \ (\sum_1^n |\beta_j|^2)^{-1/2} \sum_1^n |\beta_j|^2 \ = \ (\sum_1^n |\beta_j|^2)^{1/2} \ \leq \ \|\lambda\| \ < \ \infty \end{aligned}$$

for every sufficiently large n. Since the right side of this equation is independent of n we may take the limit as $n \to \infty$ on both sides and conclude that $\sum_1^\infty |\beta_j|^2 \leq \|\lambda\|^2 < \infty$. □

1.6.9 Corollary. *If H is a Hilbert space with dual H', then H' and H are isometrically isomorphic.*

1.6.10 Remark. If H is a complex Hilbert space it is often useful to characterize the *conjugate-linear* functionals, $\lambda : H \to \mathbb{C}$ satisfying $\lambda(\alpha x) = \overline{\alpha}\lambda(x)$ for $x \in H$ and $\alpha \in \mathbb{C}$. Minor modifications in the proof of the Riesz theorem show that there is a unique $y \in H$ such that

$$\langle y, x\rangle \ = \ \lambda(x) \qquad \text{for every} \quad x \in H \tag{1.6.2}$$

where the inner product is defined so that $\langle y, \alpha x\rangle = \overline{\alpha}\langle y, x\rangle$. In this case we also have $\|y\| = \|\lambda\|$. (Cf. Exercise 1.6.21.)

Now let's look at some examples where the Riesz theorem can be useful. We begin with the simplest.

1.6.11 Example. Let Q be an $n \times n$ matrix which is Hermitian ($Q^* = Q$) and strictly positive definite on $\mathbb{C}^n$ ($x^*Qx > 0$ for all $x \neq 0$ in $\mathbb{C}^n$). Then $(x, y) \mapsto x^*Qy$ is an inner product on $\mathbb{C}^n$ by Exercise 1.1.36.

Now given $y \in \mathbb{C}^n$ we consider the problem of solving the linear system $Qx = y$ for x. By the Riesz theorem, since $z \mapsto y^*z$ defines a bounded linear functional on $\mathbb{C}^n$, there is a unique $x \in \mathbb{C}^n$ such that $x^*Qz = y^*z$ for all $z \in \mathbb{C}^n$. This means $x^*Q = y^*$, or $Qx = y$ (since $Q^* = Q$).

1.6.12 Example. Let $\Omega \subset \mathbb{R}^d$ be open and $k(x, y) \in L^2(\Omega \times \Omega)$ be Hermitian and positive in the sense that $k(y, x) = \overline{k(x, y)}$ and

$$\int \int_{\Omega \times \Omega} k(x, y)\, \overline{u(x)}\, u(y)\, dx\, dy \;\geq\; 0$$

for all $u \in L^2(\Omega)$. Define $K : L^2(\Omega) \to L^2(\Omega)$ by $Ku(x) = \int_\Omega k(x, y)u(y)\, dy$ (cf. Example 1.4.4). We consider the problem of solving the integral equation

$$u(x) + Ku(x) \;=\; f(x) \tag{1.6.3}$$

for u when $f \in L^2(\Omega)$. Under the assumptions on k, the sesqui-linear function

$$\langle v, u\rangle \;=\; \int_\Omega v(x)\overline{u(x)}\, dx + \int \int_{\Omega \times \Omega} k(y, x)v(x)\overline{u(y)}\, dx\, dy$$

is an inner product on $L^2(\Omega)$ (cf. Exercise 1.6.18). And $\lambda(v) = \int_\Omega \overline{f(x)}v(x)\, dx$ is a bounded linear functional on $L^2(\Omega)$ (cf. Example 1.4.8). So the Riesz theorem implies there is a unique $u \in L^2(\Omega)$ such that $\langle v, u\rangle = \int \bar{f}v\, dx$, or

$$\int \Big\{\overline{u(x)} + \int \overline{k(x, y)}\, \overline{u(y)}\, dy\Big\}v(x)\, dx \;=\; \int \overline{f(x)}v(x)\, dx$$

for all $v \in L^2(\Omega)$. Thus, u satisfies (1.6.3).

1.6.13 Remark. If $w_j(x)$, $j = 1, 2, ...$ is a countable set of orthonormal functions in $L^2(\Omega)$, functions of the form

$$k(x, y) \;=\; \sum_j \lambda_j w_j(x)\overline{w_j(y)} \tag{1.6.4}$$

satify the two conditions of Example 1.6.12 when $\lambda_j \geq 0$ for all j and $\sum_j \lambda_j^2 < \infty$.

If $w_1(x), ..., w_n(x)$ is a finite set of linearly independent functions in $L^2(\Omega)$, functions of the form

$$k(x, y) \;=\; \sum_{i=1}^n \sum_{j=1}^n a_{ij} w_i(x)\overline{w_j(y)} \tag{1.6.5}$$

also satify these conditions when $A = [a_{ij}]$ is the inverse of the Gram matrix $[\int w_i(x)\overline{w_j(x)}\, dx]$ (see Exercise 1.6.17). This kernel leads to an integral operator which is the projection onto the n-dimensional subspace spanned by the w_j's.

1.6.14 Remark. We point out that if k has an eigen-function expansion of the form (1.6.4) then an explicit formula for u can be given. Assume the set w_j is countably infinite and spans $L^2(\Omega)$ (they can always be extended as such in principle). Then the identity linear transformation on $L^2(\Omega)$ has the kernel

$$\delta(x - y) \;=\; \sum_1^\infty w_j(x)\, \overline{w_j(y)}$$

in the sense that $f(x) = \sum_1^\infty w_j(x) \int \overline{w_j(y)}f(y)\, dy$ for all $f \in L^2(\Omega)$. (This is the Fourier expansion of f in the basis w_j.) Then $(I + K)u = f$ can be written

$$\sum_1^\infty (1 + \lambda_j) \Big[\int u(y)\overline{w_j(y)}\, dy\Big]\, w_j(x) \;=\; \sum_1^\infty \Big[\int f(y)\overline{w_j(y)}\, dy\Big]\, w_j(x)\,.$$

Thus we can algebraically solve for the Fourier coeffieicnts $\int u(y)\overline{w_j(y)}\, dy$ of u.

1.6.15 Example. Let $f \in L^2(-1,1)$. The linear functional $f : v \mapsto \int_{-1}^1 f(x)v(x)\,dx$ is bounded on $H_0^1(-1,1)$. For

$$|\int_{-1}^1 f(x)v(x)\,dx| \le (\int_{-1}^1 |f(x)|^2\,dx)^{1/2}\,(\int_{-1}^1 |v(x)|^2\,dx)^{1/2} \le (\int_{-1}^1 |f(x)|^2\,dx)^{1/2}\,\|v\|_1\ .$$

There is therefore a (unique) $u \in H_0^1(-1,1)$ such that

$$\int_{-1}^1 u(x)v(x) + u'(x)v'(x)\,dx \;=\; \int_{-1}^1 f(x)v(x)\,dx\ .$$

If the solution $u \in H_0^1(-1,1)$ is sufficiently smooth the left side of this equation equals $\int_{-1}^1 (u(x) - u''(x))v(x)\,dx$. This suggests that u may be, in some sense, a solution of the boundary value problem

$$u(x) - u''(x) \;=\; f(x)$$

with $u(-1) = u(1) = 0$. Later we will say much more about this.

If T is a symmetric (Hermitian) positive definite matrix, the equation $Tx = y$ can always be solved for x when we are given $y \in \mathbb{R}^n$ ($\mathbb{C}^n$). In Exercise 1.6.22 you are asked to prove the infinite dimensional version of this, and show how the Riesz theorem can be used to solve such linear equations. *Two different inner products will be used simultaneously.*

1.6.16 Exercise. Let δ be the evaluation linear functional on $H_0^1(-1,1)$, that is, $\delta(v) = v(0)$. We have seen that δ is bounded on $H_0^1(-1,1)$. Show that there is a unique $u \in H_0^1(-1,1)$ such that

$$\int_{-1}^1 u(x)v(x) + u'(x)v'(x)\,dx \;=\; v(0)\ .$$

1.6.17 Exercise. Let $k(x,y)$ be as in (1.6.5). Show that $\int\int_{\Omega\times\Omega} k(x,y)\overline{u(x)}u(y)\,dx\,dy \ge 0$ for all $u \in L^2(\Omega)$.

1.6.18 Exercise. Let $\lambda > 0$, perhaps small, and $k(x,y)$ and K be as in Example 1.6.12. Show that

$$\langle v, u\rangle \;=\; \int_\Omega v(x)\Big\{\lambda\overline{u(x)} + \int_\Omega k(y,x)\,\overline{u(y)}\,dy\Big\}\,dx \tag{1.6.6}$$

is an inner product on $L^2(\Omega)$ which is equivalent to the usual one, $(v,u) \mapsto \int v\overline{u}\,dx$.

1.6.19 Exercise. Let $\lambda > 0$, perhaps small, and $k(x,y)$ and K be as in Example 1.6.12. Use the inner product (1.6.6) on $L^2(\Omega)$ and the Riesz theorem to show that the integral equation $(\lambda + K)u = f$ has a unique solution $u \in L^2(\Omega)$ given any $f \in L^2(\Omega)$.

1.6.20 Exercise. Prove that $H^1(a,b) \subset C(a,b)$. This fact is related to the delta function being continuous. Hint: If $u \in H^1(a,b)$, use the estimate in Example 1.4.7 to bound $|u(x) - u(y)|$ when $x, y \in (a,b)$. Hence show that $u(y) \to u(x)$ as $y \to x$.

1.6.21 Exercise. Prove Remark 1.6.10.

1.6.22 Exercise. Let X be a Hilbert space and $T : X \to X$ be a bounded linear transformation satisfying $\langle Tx, z\rangle = \langle x, Tz\rangle$ for all $x, z \in X$, and $\langle Tx, x\rangle \ge c\,\|x\|^2$ for all $x \in X$ where $c > 0$ is some constant. Let $y \in X$ be any element of X. Show that there exists a unique $x \in X$ that satisfies

$$\langle Tx, z\rangle = \langle y, z\rangle \tag{1.6.7}$$

for all $z \in X$. Show that this x is a solution to the equation

$$Tx = y\ . \tag{1.6.8}$$

1.7 Contraction Mappings

The *contraction mapping theorem* has several important applications in analysis.

1.7.1 Definition. Let X be a metric space with the distance between $x, y \in X$ denoted by $d(x, y)$. A function $f : X \to X$ is a *contraction mapping* on X if there exists an $\alpha \in [0, 1)$ such that

$$d(f(x), f(y)) \leq \alpha\, d(x, y)$$

for all $x, y \in X$.

Every contraction mapping is uniformaly continuous on X (Exercise 1.7.6).

1.7.2 Theorem (contraction mapping)**.** *Let X be a complete metric space and $f : X \to X$ a contraction mapping. Then f has a unique fixed point; i.e., there is a unique $x \in X$ such that $f(x) = x$.*

Proof. Uniqueness. Suppose $f(x) = x$ and $f(y) = y$. Since f is a contraction

$$d(x, y) = d(f(x), f(y)) \leq \alpha\, d(x, y) < d(x, y)$$

which is a contradiction unless $d(x, y) = 0$.

Existence. In the proof we construct a sequence x_n which converges to the fixed point x of f. Begin with any $x_0 \in X$. Successively define

$$x_{n+1} = f(x_n) \qquad \text{for} \quad n \in \mathbb{N}_0.$$

Observe that $d(x_2, x_1) = d(f(x_1), f(x_0)) \leq \alpha\, d(x_1, x_0)$. Similarly $d(x_3, x_2) = d(f(x_2), f(x_1)) \leq \alpha\, d(x_2, x_1)$, and combining these two relations gives $d(x_3, x_2) \leq \alpha^2\, d(x_1, x_0)$. An easy induction arguement shows that

$$d(x_{n+1}, x_n) \leq \alpha^n\, d(x_1, x_0) \tag{1.7.1}$$

for all $n \in \mathbb{N}$.

Now this inequality shows that x_n is a Cauchy sequence. For if $m > n$ in $\mathbb{N}$ the triangle inequality gives

$$\begin{aligned} d(x_m, x_n) \;&\leq\; \sum_{j=n}^{m-1} d(x_{j+1}, x_j) \;\leq\; \sum_{j=n}^{m-1} \alpha^j\, d(x_1, x_0) \\ &=\; \alpha^n \sum_{j=0}^{m-n-1} \alpha^j\, d(x_1, x_0) \;\leq\; \frac{\alpha^n}{1-\alpha} d(x_1, x_0). \end{aligned}$$

(Cf. Exercise 1.7.8.) Since $\alpha < 1$ the right hand side can be made as small as desired by taking n large enough.

Finally the completeness of X implies there is an $x \in X$ such that $\lim_{n\to\infty} x_n = x$. And since f is continuous,

$$x = \lim x_{n+1} = \lim f(x_n) = f(\lim x_n) = f(x)\,,$$

so x is a fixed point of f. □

We end this section by listing five important applications of the contraction mapping theorem.

- The *inverse function theorem* gives conditions when non-linear equations $f(x) = y$ can be solved for x, even when x and y are elements of Banach spaces. The theorem is only local, x close to x_0 and y close to y_0 when $f(x_0) = y_0$ is known, and the proof relies on the contraction mapping theorem.
- The *implicit function theorem*; its proof uses the inverse function theorem.
- Solutions of *small perturbations* of linear equations $(I + A)x = y$ when $\|A\| < 1$. (See Exercise 1.7.3.)
- Solutions of integral equations with small parameter: $u(x) + \lambda \int k(x, y)\, u(y)\, dy = f(x)$ when $\lambda > 0$ is small enough. This is a special, but important, case of the last bullet.

- The most general existence theorem we have for ordinary differential equations is based on Picard's method of successive approximations, and is proved using the contraction mapping theorem. This will be studied in Chapter 2.

The first two items remain two of our most powerful tools for solving systems of non-linear equations, even in infinite dimensions. The third item has many applications; it is useful because it turns the difficult problem of inverting a linear operator equation into the easier problem of repeated multiplication and passing to a limit. (This is not so hard in finite dimensions but much more so in infinite dimensions.) The application to integral equations is not too important because the theory of compact operators (which integral operators are) gives us a much more complete theory for the solution of these equations. And the application to ordinary differential equations is perhaps the most important result in all of ODEs. Picard's theorem remains the most important theoretical result we have, even though the algorithm is so inefficient that it cannot be used for numerical computation.

1.7.3 Exercise. Let X be a Banach space, $A \in \mathscr{B}(X)$, $b \in X$, and $f : X \to X$ be defined by $f(x) = Ax + b$. Show that f is a contraction if and only if $\|A\| < 1$.

1.7.4 Exercise. Show that the only fixed point of a *linear* contraction is $x = 0$.

1.7.5 Exercise. If $f(x) = Ax + b$ is an affine contraction as in Exercise 1.7.3, show that the fixed point x of f satisfies $(I - A)x = b$. Show that

$$x = \sum_{k=0}^{\infty} A^k\, b$$

where the series converges in X. (Cf. Theorem 1.3.19.)

1.7.6 Exercise. Let X be a metric space. Prove that every contraction $f : X \to X$ is uniformly continuous on X. Give a formula for δ, as a function of ϵ, that works in the $\epsilon\delta$-definition of continuity.

1.7.7 Exercise. Prove inequality (1.7.1).

1.7.8 Exercise. Let $0 < \alpha < 1$. Show that, for any $n \in \mathbb{N}$,

$$\sum_{j=0}^{n} \alpha^j < \frac{1}{1-\alpha}\,.$$

1.8 Real Valued Functions of a Vector Variable

In this section we assume the functions are real valued. Our goal is to develop a theory of optimization for functions defined on infinite dimensional vector spaces. Here are two example applications of this situation, and many others will be given in Chapter 4 (calculus of variations).

1.8.1 Example. Let $[a, b]$ be a closed bounded interval and $f : [a, b] \to \mathbb{R}$ be continuously differentiable on $[a, b]$. The real number

$$\ell(f) \;=\; \int_a^b \sqrt{1 + (f'(x))^2}\, dx$$

is the length of the curve in $\mathbb{R}^2$ which joins the points $(a, f(a))$ and $(b, f(b))$ and is the graph of the function f. The mapping $f \mapsto \ell(f)$ is a functional on the vector space $C^1([a, b])$.

1.8.2 Example. Consider a physical system of n particles moving in $\mathbb{R}^3$. Let particle i have mass m_i and position $x_i(t), y_i(t), z_i(t)$ at time t. The kinetic energy of this system at time t is

$$T(t) \;=\; \frac{1}{2} \sum_{i=1}^{n} m_i(\dot{x}_i(t)^2 + \dot{y}_i(t)^2 + \dot{z}_i(t)^2)\;.$$

In many situations the dynamics of this system is governed by a potential evergy function

$$U(t) \;=\; U(t, x_1(t), y_1(t), z_1(t), ..., x_n(t), y_n(t), z_n(t)) \;.$$

In this case the motion of these particles between the times t_0 and t_1 is given by the trajectory $t \mapsto (x_1(t), y_1(t), z_1(t), ..., x_n(t), y_n(t), z_n(t))$ that minimizes the *action functional*

$$\int_{t_0}^{t_1} T(t) - U(t)\, dt \;.$$

1.8.3 Definition (directional derivative or variation)**.** Let X be a normed vector spaces over $\mathbb{R}$, and let $U \subset X$ be open. A function $F : U \to \mathbb{R}$ has a *directional derivative* in the direction $h \in X$, $h \neq 0$, at the point $x \in U$ if the function $\phi(t) = F(x + th)$ of the real variable t is differentiable at $t = 0$, that is, if

$$\delta F(x; h) \stackrel{def}{=} \frac{d\phi}{dt}(0) = \lim_{t \to 0} \left[\frac{F(x + th) - F(x)}{t} \right]$$

exists. $\delta F(x; h)$ is called the *directional derivative* (or the *Gateaux derivative* or *variation*) of F at x in the direction h.

1.8.4 Example. Let $F : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ where Ω is open. For $x \in \Omega$ the directional derivative of F at x, in the direction $h \in \mathbb{R}^d$, if it exists, is the function

$$\delta F(x; h) = \sum_{j=1}^{n} \frac{\partial F}{\partial x_j}(x)\, h_j$$

each component being computed from the definition by the ordinary chain rule on $\mathbb{R}^d$. The Gateaux derivative exists if each of the n partial derivatives of F exist at x.

1.8.5 Example. Let $X = C^1(0, 1) \cap H^1(0, 1)$ denote the vector space of real-valued functions, and $f : X \to \mathbb{R}$ be given by

$$f(u) = \int_0^1 u(s)^2 + u'(s)^2\, ds.$$

Set

$$\begin{aligned} \Phi(t) \;&=\; f(u + th) \;=\; \int_0^1 (u(s) + th(s))^2 + (u'(s) + th'(s))^2\, ds \\ &= \int_0^1 u(s)^2 + 2tu(s)h(s) + t^2 h(s)^2 + u'(s)^2 + 2tu'(s)h'(s) + t^2 h'(s)^2\, ds \end{aligned}$$

where $u, h \in X$ and $t \in \mathbb{R}$. Using the linearity of the integral with respect to s we calculate

$$\frac{d}{dt}\Phi(t) = 2\int_0^1 u(s)h(s)\, ds + 2t\int_0^1 h(s)^2\, ds + 2\int_0^1 u'(s)h'(s)\, ds + 2t\int_0^1 h'(s)^2\, ds$$

and

$$\frac{d}{dt}\Phi(0) = 2\int_0^1 u(s)h(s) + u'(s)h'(s)\, ds.$$

This gives us a formula for the Gateaux derivative of f at $u \in X$.

If u also happens to be in $C_b^2(0, 1)$ (which is a sub-vector space of X), we can integrate the second term by parts one time to get

$$\frac{d}{dt}\Phi(0) = 2\int_0^1 [u(s) - u''(s)]h(s)\, ds + u'(1)h(1) - u'(0)h(0)$$

which is an expression without any derivatives on the function h. Later we will see the value of such expressions.

1.8.6 Example. Let $C_b(0,1)$ denote the vector space of continuous and bounded functions on the open interval $(0,1)$, and let $f : (0,1) \times \mathbb{R} \to \mathbb{R}$ be continuously differentiable and satisfy

$$|f(t,x)| \leq M\,|x|$$

for all $x \in \mathbb{R}$ and $t \in (0,1)$, where $M > 0$ is a constant independent of t. Define $F : C_b(0,1) \to C_b(0,1)$ by

$$F(u)(t) = \int_0^t f(s,u(s))\,ds.$$

Let's calculate the variation of F in the direction of $h \in C_b(0,1)$.

$$\begin{aligned}\delta F(u;h)(t) &= \frac{d}{d\tau}|_{\tau=0} F(u+\tau h)(t)\\ &= \int_0^t \frac{d}{d\tau}|_{\tau=0} f(s,u(s)+\tau h(s))\,ds\\ &= \int_0^t \frac{\partial f}{\partial u}(s,u(s))h(s)\,ds.\end{aligned}$$

Here we can differentiate under the integral sign because the differentiated integrand is continuous and remains bounded. The resulting function $\delta F(u;h)$ on $(0,1)$ is bounded and continuous.

Local Optimization

1.8.7 Definition. Let X be a normed vector space and $f : V \subset X \to \mathbb{R}$ where V is open. We say that f has a *local minimum (maximum)* at $x \in V$ if there is a neighborhood U of x which is contained in V and on which f satisfies $f(x) < f(y)$ $(f(x) > f(y))$ for all $y \in U$ with $y \neq x$.

f is said to have a *global minimum (maximum)* at x on V if f satisfies $f(x) < f(y)$ $(f(x) > f(y))$ for all $y \in V$ with $y \neq x$.

We say that f has an *extremal* at x if f has either a maximum or a minimum at x.

A point $x \in V$ at which $\delta f(x)$ exists is called a *stationary* or *critical point* if $\delta f(x;h) = 0$ for all $h \in X$.

1.8.8 Theorem (first derivative test)**.** *Let X be a normed vector space and $U \subset X$ an open subset. If the functional $f : U \to \mathbb{R}$ has a Gateaux derivative at $x \in U$, and if f has a local extremal at x, then*

$$\delta f(x;h) = 0 \qquad \text{for all} \quad h \in X\ .$$

Proof. Here is the argument when f has a minimum. For every $h \neq 0$ in X and real $t \neq 0$ sufficiently small, $f(x+th) > f(x)$. Thus, if $t > 0$, $\lim_{t\downarrow 0}(f(x+th)-f(x))/t \geq 0$. Similarly, if $t < 0$, $\lim_{t\uparrow 0}(f(x+th)-f(x))/t \leq 0$. (In the second case the numerator is positive but the denominator is negative as $t \uparrow 0$.) Thus, $\delta f(x;h) = 0$. □

1.8.9 Exercise. Let $C_b([0,\infty))$ denote the vector space of bounded real-valued functions on $[0,\infty)$, and let $r > 0$ and $f \in C_b([0,\infty))$ be given. Define $F : C_b([0,\infty)) \to \mathbb{R}$ by

$$F(u) \;=\; \int_0^\infty (u(s) - f(s))^2\, e^{-rs}\,ds.$$

Calculate the variation of F.

1.8.10 Exercise. Let $C_b([0,\infty))$, r, and f be as in the last exercise. Define $F : C_b([0,\infty)) \to \mathbb{R}$ by

$$F(u) \;=\; \int_0^\infty u(s) f(s)\, e^{-rs}\,ds.$$

Calculate the variation of F.

1.9 Compact and Self-Adjoint Operators

Dettman, section 2.8, has an alternate, good, introduction to compact operators.

If $T : \mathbb{K}^n \to \mathbb{K}^m$ is linear then $\dim \mathscr{R}(T) + \dim \mathscr{N}(T) = n$. When $m = n$ we can conclude that T is onto just by knowing it is one-to-one. When X is an infinite dimensional Banach space and $T \in \mathscr{B}(X)$ the Fredholm alternative provides a similar condition if T has the form $T = I + K$ where K is 'compact' (to be defined shortly).

Let X and Y be Banach spaces. Recall that the (vector) space of bounded (continuous) linear transformations from X to Y is denoted $\mathscr{B}(X, Y)$. This is the set of $T : X \to Y$ for which the bound $\|Tx\| \leq C\|x\|$ holds for all $x \in X$, and where $C > 0$ is independent of x. The smallest such C is the operator norm of T and is denoted $\|T\|$.

1.9.1 Definition. Let $B = \{x \in X \ ; \ \|x\| < 1\}$ be the open unit ball in X, $T \in \mathscr{B}(X, Y)$, and $\overline{T(B)}$ the closure of the image of B under T. The linear transformation T is *compact*, or *completely continuous*, if $\overline{T(B)}$ is a compact subset of Y.

1.9.2 Remark. The property that $\overline{T(B)}$ be compact in Y is equivelant to either of the following conditions:

- $\overline{T(\overline{B})}$ is compact where $\overline{B} = \{x \in X \ ; \ \|x\| \leq 1\}$;
- every sequence y_n in $T(B)$ has a limit point in Y. This means that there is a sub-sequence y_{n_k} such that $y = \lim_k y_{n_k}$ exists in Y.
- every bounded sequence x_n in X gets mapped, by T, to a sequence $y_n = Tx_n$ in Y which has a limit point in Y.

If Y is finite dimensional, every linear transformation T is compact since $\overline{T(B)}$ is closed and bounded. Example 1.3.18(g) provides examples of compact operators when Y is infinite dimensional.

Integral operators and inverses of differential operators are often compact. This is why the theory of these operators is important in applications.

We can now lists some of the basic properties of compact operators. Many of these properties have to do with eigenvalues which may be complex, so it seems natural to adopt *complex* Banach spaces as our setting.[8]

1.9.3 Proposition. *Let X and Y be complex Banach spaces and $\lambda \in \mathbb{C}$.*

(a) If $T \in \mathscr{B}(X, Y)$ has finite rank then T is compact.

(b) If $T \in \mathscr{B}(X)$ is compact and $\lambda \neq 0$ then $\dim \mathscr{N}(T - \lambda I) < \infty$.

(c) If $\dim X = \infty$ and $T \in \mathscr{B}(X)$ is compact, then T is not onto.

(d) If X, Y, and Z are Banach spaces, $S \in \mathscr{B}(X, Y)$, and $T \in \mathscr{B}(Y, Z)$, then TS is compact if either *S or T is compact.*

(e) The set of compact operators on X is a two-sided ideal in the ring $\mathscr{B}(X)$. In particular, if S is compact and T is bounded, both ST and TS are compact.

(f) If T_n is a sequence of compact operators in $\mathscr{B}(X, Y)$ and $\|T - T_n\| \to 0$ as $n \to \infty$ for some $T \in \mathscr{B}(X, Y)$, then T is compact.

(g) $T \in \mathscr{B}(X, Y)$ is compact if and only if the transpose $T' \in \mathscr{B}(Y', X')$ is compact.

(h) If X and Y are Hilbert spaces then $T \in \mathscr{B}(X, Y)$ is compact if and only if there is a sequence T_n of finite rank operators in $\mathscr{B}(X, Y)$ such that

$$\|T - T_n\| \to 0 \qquad \text{as } n \to \infty.$$

Part (h) is not necessarily true when X and Y are Banach spaces.

The set $\mathscr{N}(T - \lambda I)$ is the eigen-space of T corresponding to the eigen-value λ.

[8] This issue arises when studying matrices on finite dimensional vector spaces as well.

Fredholm Alternative

1.9.4 Theorem (Fredholm alternative)**.** *Let X be a complex Banach space and $T \in \mathscr{B}(X)$ be compact. For $\lambda \in \mathbb{C}$ define the subspaces*

$$V_\lambda = \mathscr{N}(T - \lambda I) \subset X \quad \text{and} \quad V'_\lambda = \mathscr{N}(T' - \lambda I) \subset X'.$$

Then the following holds.

a) $V_\lambda = \{0\}$ except for a finite or countably infinite subset of $\lambda \in \mathbb{C}$. If this exceptional set is infinite, it is bounded and its only point of accumulation is 0.

b) If $\lambda \neq 0$ then $\dim V_\lambda < \infty$.

c) If $\lambda \neq 0$ then $\dim V_\lambda = \dim V'_\lambda$.[9]

d) If $\lambda \neq 0$ then $\mathscr{R}(T - \lambda I)$ is closed.

e) If $\lambda \neq 0$ then, $(\lambda I - T)u = f$ has a solution if and only if $f \perp V'_\lambda$.[10]

f) If $\lambda \neq 0$ then, $\lambda I - T$ is onto if and only if it is one-to-one.

Here, we have used I to denote the identity operator on both X and X'.

In the preceeding statements, notice that the operator $I - T$ is qualitatively the same as the operator $\lambda I - T$ if T is compact; we have only to interchange T and T/λ which is still compact with the same domain and range (but now with different eigenvalues). Thus, the theorems can often be stated in two equivalent ways, using either of these two operators.

Proof. See Folland, section 0.F, for a lucid proof when X is a Hilbert space (the most important case). □

Self-Adjointness Recall that an operator $T \in \mathscr{B}(H)$ on a Hilbert space H is called *self-adjoint* if $\langle Tu, v\rangle = \langle u, Tv\rangle$ for all $u, v \in H$. If T is a self-adjoint compact operator on a complex Hilbert space H, an orthonormal basis for H can be constructed from the the eigenvectors of T.

1.9.5 Theorem (spectral theorem)**.** *Let H be a separable complex Hilbert space and $T \in \mathscr{B}(H)$ be compact and self-adjoint. Then the eigen-values of T are real, and the eigen-vectors of T form a complete orthonormal basis for H. This means that there is a countable, or finite if $\dim(H) < \infty$, set $\{e_j \,;\, j \in \mathbb{N}\}$ with the following properties.*

a) $\langle e_i, e_j\rangle = \delta_{ij}$ for all $i, j \in \mathbb{N}$.

b) $Te_j = \lambda_j e_j$ for all $j \in \mathbb{N}$.

c) For any $u \in H$, $u = \sum_{j=1}^{\infty} \alpha_j e_j$, convergent in H, where $\alpha_j = \langle e_j, u\rangle$ and $\sum_1^\infty |\alpha_j|^2 < \infty$.

d) If u is given the preceding expansion then $Tu = \sum_{j=1}^{\infty} \lambda_j \alpha_j e_j$. This series converges in the norm of H.

Part (d) can be expressed by saying that T has the expansion

$$T \;=\; \sum_{j=1}^{\infty} \lambda_j \, e_j \, e_j^* \tag{1.9.1}$$

where e_j^* is simply notation that is ment to suggest the conjugate transpose of the 'column vector' e_j so that $e_j^* u = \langle e_j, u\rangle$. Thus e_j^* is the bounded linear functional (element of the dual of H) given by $e_j^* : u \mapsto \langle e_j, u\rangle$, the complex number α_j. So $e_j\, e_j^*$ is a rank one linear transformation of H into itself, which is the projection onto the span of e_j. One sometimes refers to (1.9.1) as an expansion of T in a *weighted sum of projections.*

The following lemma is used in the proof, and is interesting in its own right.

1.9.6 Lemma. *If H is a complex Hilbert space and $T \in \mathscr{B}(H)$ is self-adjoint, then all eigenvalues of T are real. And if λ and μ are distinct eigenvalues of T with corresponding eigenvectors $u, v \in H$, then $u \perp v$.*

This result says that for distinct eigenvalues of T, the eigen-subspaces in H are orthogonal to each other.

[9] If *conjugate* bi-linear duality is used, as is often the case when X is a complex Hilbert space, the correct statement is $\dim V_\lambda = \dim V'_{\overline{\lambda}}$.

[10] If conjugate duality pairing is used, replace V'_λ by $V'_{\overline{\lambda}}$.

Proof. Let λ satisfy $Tu = \lambda u$ for some $u \neq 0$ in H. Then

$$\lambda \langle u, u \rangle \;=\; \langle \lambda u, u \rangle \;=\; \langle Tu, u \rangle \;=\; \langle u, Tu \rangle \;=\; \langle u, \lambda u \rangle \;=\; \overline{\lambda} \langle u, u \rangle \;.$$

This shows that $\lambda = \overline{\lambda}$ if $\|u\| > 0$.

If λ and μ are distinct (real) eigenvalues of T and u and v are (non-zero) eigenvectors corresponding to them, then we may write

$$\lambda \langle u, v \rangle \;=\; \langle \lambda u, v \rangle \;=\; \langle Tu, v \rangle \;=\; \langle u, Tv \rangle \;=\; \langle u, \mu v \rangle \;=\; \mu \langle u, v \rangle \;.$$

This shows $\langle u, v \rangle = 0$ if $\lambda \neq \mu$. □

The Rayleigh-Ritz procedure is a classical method for constructing eigenvectors of compact operators. See Dettman, page 94.

1.9.7 Example. Suppose T is a self-adjoint, compact operator on a Hilbert space X, and suppose we know that λ is not an eigenvalue of T. Then, given any $f \in X$, we may solve the equation

$$(\lambda I - T)u \;=\; f$$

explicitly for u using the eigenvectors of T.

Let $T\, e_j = \lambda_j\, e_j$ for $j \in \mathbb{N}$, where the e_j's form a complete orthonormal set for X and the λ_j's need not all be distinct. We may expand $f = \sum_1^\infty \beta_j\, e_j$ (convergent in X[11]) where $\beta_j = \langle e_j, f \rangle$, and similarly $u = \sum_1^\infty \alpha_j\, e_j$ but with the coefficients α_j unknown. Since T is continuous and linear

$$(\lambda I - T)u \;=\; \sum_{j=1}^{\infty} \alpha_j\, (\lambda I - T) e_j \;=\; \sum_{j=1}^{\infty} \alpha_j\, (\lambda - \lambda_j)\, e_j \;.$$

Now if we set this expression equal to $\sum_1^\infty \beta_j\, e_j$, and use the linear independence of the e_j's, we have $\alpha_j\, (\lambda - \lambda_j) = \beta_j$ which can be solved for α_j in terms of known quantities. Thus, the solution is

$$u \;=\; \sum_{j=1}^{\infty} \frac{\beta_j}{\lambda - \lambda_j}\, e_j \;,$$

and this series is convergent in X.

The following theorem is a corollary of the spectral theorem and gives a very important application to differential operators. We want to give a condition when non-compact, even unbounded, self-adjoint operators also have an expansion of the form (1.9.1). First we need a definition.

1.9.8 Definition. Let X be a Banach space and $\mathcal{D} \subset X$ be a subspace. A linear transformation $L : \mathcal{D} \to X$ is called a *linear operator* on X if $\mathcal{D}$ is dense in X. The subspace $\mathcal{D}$ is call the domain of L. We usually write $\mathcal{D}_L$ to indicate the dependence of the subspace on L. If X is a Hilbert space, we say L is *self-adjoint* if

$$\langle Lu, v \rangle \;=\; \langle u, Lv \rangle$$

for all $u, v \in \mathcal{D}_L$.

This definition is a slight generalization of Definition 1.3.1 in that the domain $\mathcal{D}_L$ need not be all of X. This distinction is only significant when L is unbounded; when L is bounded Proposition 1.3.14 shows that L can be extended to all of X. Here are a couple of typical examples.

1.9.9 Example. Let $\Omega \subset \mathbb{R}^d$ be open and $L = \sum_{|\alpha|=0}^m a_\alpha(x)\, \partial^\alpha$ be a partial differential operator on Ω with coefficient function $a_\alpha \in C^\infty(\bar{\Omega})$. Then L is a linear operator on $L^2(\Omega)$ if we define its domain as $\mathcal{D}_L = \{u \in L^2(\Omega) \;;\; Lu \in L^2(\Omega)\}$. To see that $\mathcal{D}_L$ is dense in $L^2(\Omega)$ we simply notice that $\mathcal{D}_L$ certainly contains $C^\infty(\bar{\Omega})$ which is itself a dense subset of $L^2(\Omega)$.

[11] TO STEVE: At this point I expect the reader to have little understanding of this 'generalized' Fourier series expansion; I either need to add some details or preceed by something on Fourier series. I could cite an earlier theorem that every separable H-sp is isomorphic to ℓ^2

1.9.10 Example. Consider Example 1.3.18. If $t_j \to \infty$ the operator T is not bounded. But we can define $\mathcal{D}_T = D$ where the set D is given in that example. The vector subspace consisting of all sequences $x = (a_1, a_2, ...) \in \ell^2(\mathbb{N})$ such that only a finite number of the a_j's are not zero is both contained in D and dense in $\ell^2(\mathbb{N})$.

In the following theorem it becomes apparent why we require $\mathcal{D}_L$ to be a dense subspace.

1.9.11 Theorem. *Let H be a separable Hilbert space and $L : H \to H$ a self-adjoint linear transformation which need not be bounded, nor even defined on all of H. Assume that there is a $\lambda \in \mathbb{C}$ such that the resolvent*

$$T \;=\; (\lambda I - L)^{-1}$$

is compact and self-adjoint. Then there is an orthonormal basis $\{e_j \,;\, j \in \mathbb{N}\}$ for H and a sequence of complex numbers λ_j with the following properties.

a) $Le_j = \lambda_j\, e_j$ for all $j \in \mathbb{N}$.

b) $|\lambda_j| \to \infty$ as $j \to \infty$, and L is necessarily unbounded on H.

c) The domain of L is precisely the set

$$\{u = \sum_1^\infty \alpha_j\, e_j \in H \,;\, \sum_1^\infty |\lambda_j|^2\, |\alpha_j|^2 < \infty\} \;.$$

d) L has an expansion of the form

$$L \;=\; \sum_{j=1}^\infty \lambda_j\, e_j\, e_j^*$$

by which we mean that, for every $u = \sum_1^\infty \alpha_j\, e_j$ in the domain of L, we have $Lu = \sum_1^\infty \lambda_j\, \alpha_j\, e_j$.

Proof Sketch. By the spectral theorem there exists an orthonormal basis e_j for H which consists of eigenvectors of T with eigenvalues μ_j, a sequence in $\mathbb{C}$ converging to zero. Set

$$T \;=\; (\lambda I - L)^{-1} \;=\; \sum_1^\infty \mu_j\, e_j\, e_j^* \;.$$

Now $T^{-1} = \sum_1^\infty \mu_j^{-1}\, e_j\, e_j^*$ is unbounded on H since $\mu_j^{-1} \to \infty$ as $j \to \infty$. But this expansion is still valid for any $u = \sum_1^\infty \alpha_j\, e_j$ in H for which $T^{-1}u = \sum_1^\infty \mu_j^{-1}\, e_j\, \langle e_j u\rangle$ also lies in H. That is, whenever $\sum_1^\infty |\mu_j^{-1}|^2\, |\alpha_j|^2 < \infty$. So we have $\lambda I - L = \sum_1^\infty \mu_j^{-1}\, e_j\, e_j^*$ when this series, upon application to some $u \in H$, converges in H. But it is clear that $I = \sum_1^\infty e_j\, e_j^*$ (all eigenvalues equal 1). So we also have $L = \sum_1^\infty (1 - \mu_j^{-1})\, e_j\, e_j^*$. And the theorem holds with $\lambda_j = 1 - \mu_j^{-1}$. □

Many of the special functions of mathematical physics arise as eigenfunctions of a self-adjoint differential operator L, for instance a Sturm-Liouville operator, on some Hilbert space.

1.9.12 Exercise. Show that if $T : X \to Y$ and $S : Y \to Z$ are (linear and) bounded, and if either S *or* T is compact, then ST is compact.

1.9.13 Exercise. Prove Remark 1.9.2.

1.9.14 Exercise. Define the partial sums of (1.9.1) by $T_n = \sum_{j=1}^n \lambda_j\, e_j\, e_j^*$. Show that $\|T - T_n\| \le \max\{|\lambda_j| \,;\, j \ge n+1\}$, so $\|T - T_n\| \to 0$ as $n \to \infty$ and (1.9.1) converges in the operator norm.

2 Initial Value Problems for Ordinary Differential Equations

Most of this chapter is concerned with ordinary differential equations of the form

$$\dot{x}(t) \;=\; f(t, x(t)) \tag{2.0.1}$$

where $t \in \mathbb{R}$, $\dot{} = \frac{d}{dt}$, and the function x takes values in $\mathbb{R}^n$. Conditions on $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ will be given to ensure equation (2.0.1) has a unique solution, provided that an initial condition

$$x(t_0) = x_0 \tag{2.0.2}$$

is specified for some $t_0 \in \mathbb{R}$ and some $x_0 \in \mathbb{R}^n$.

For linear constant coefficient equations of the form

$$\dot{x}(t) \;=\; A\,x(t) + f(t)\;, \tag{2.0.3}$$

where A is an $n \times n$ matrix and f is an n-vector valued function, an explicit formula for the solution can be given in terms of the matrix exponential

$$e^{At} \;=\; \sum_{0}^{\infty} (At)^k/k!\;.$$

When the A and f in (2.0.3) depend analytically on t in a neighborhood of t_0, (2.0.3) has the form

$$\dot{x}(t) \;=\; A(t)\,x(t) + f(t) \tag{2.0.4}$$

and this linear variable coefficient equation can be solved by substituting a power series

$$x(t) \;=\; \sum_{k=0}^{\infty} a_k\, t^k$$

and recursively solving for the unknown n-vector coefficients a_k. In general the existence of solutions for equation (2.0.1), and even (2.0.4), must be proved using the method of successive approximations and the contraction mapping theorem.

All the above methods are constructive, that is, they can in principle be used to compute approximations to the solution $x(t)$. However numerical methods, such as Runge-Kutta, are much more efficient; difficult differential equations today are solved using efficient numerical methods backed by good theoretical understanding.

Examples of Important ODEs

- Exponential grow (bacteria in a petri dish, humans on earth, money in a bank account with interest compounded continuously):

  $$\dot{y} \;=\; ry$$

  where $r > 0$ is the growth rate.

- Exponential decay (radioactive decay): $\dot{y} \;=\; -ry$ where $r > 0$ is the decay rate. The time $T = \ln 2/r$ is the half-life.

- Financial discount rates, the value of money decreasing over time, may be modeled by the equation

  $$\dot{y} \;=\; -r(t)\,y$$

  where $r(t)$ is the discount rate at time t.

- The Solow-Swan model of exogenous economic growth is

$$\dot{k}(t) \;=\; s\,k(t)^{\alpha} - (n+g+\delta)k(t)$$

where $k(t)$ is the capital intensity, s is the share of capital saved for investment, α is the output elasticity, n is the number of workers growth rate, g is the technology growth rate, and δ is the stoke depreciation rate.

- The Sethi model describing the growth of sales over time in response to advertising is

$$\dot{s}(t) \;=\; r\,u(t)\,\sqrt{1-s(t)} - \delta\,s(t)$$

where $s(t)$ is the market share (to be found), $u(t)$ is the advertising rate (under our control), r is the coefficient of advertising effectiveness (assumed constant), and δ is a decay constant.

- An object in free-fall through the earth's atmosphere has velocity $v(t)$ which satisfies the differential equation

$$m\dot{v} = mg - rv$$

if its change in altitude is small enough that the acceleration g due to gravity is constant, and where its coefficient of air resistance is r. The asymptote $v_0 = \lim_{t\to\infty} v(t) = mg/r$ is called the terminal velocity. (This velocity satisfies the equation $\dot{v} = 0$.)

- Non-linear first order equations: Population growth with limited carrying capacity:

$$\dot{y} \;=\; ry(k-y)$$

where r is the growth rate and k is the carrying capacity of the ecosystem.

- Two competing populations, such as predator-prey or two opposing armies, can be modeled by the Lotka-Volterra equations

$$\begin{pmatrix}\dot{x}(t)\\ \dot{y}(t)\end{pmatrix} \;=\; \begin{pmatrix}a\,x(t) + b\,x(t)\,y(t)\\ -c\,x(t)\,y(t) + d\,y(t)\end{pmatrix}$$

where the real coefficients a, b, c, d are all positive.

- Simple linear mechanics: A particle with mass m moving under a force propotional to its position (e.g., hooke's law), and subject to friction, obeys an equation of the form

$$m\,\ddot{x}(t) + r\,\dot{x}(t) + k\,x(t) \;=\; f(t)$$

where r is the resistence or drag, k is (for instance) the spring constant or other proportionality factor, and f represents any other forces on the particle.

- Simple circuit: A simple LRC circuit (composed of inductors, resistors, and capacitors) obeys an equation of the form

$$L\,\ddot{I}(t) + R\,\dot{I}(t) + \frac{1}{C}\,I(t) \;=\; \dot{V}(t)$$

where $I(t)$ is the current in the wire, L the inductance, R the resistance, C the capacitance, and $\dot{V}$ the time derivative of the impressed voltage. (In your home the voltage is a 'sinusoid' driven at 60 cycles per second (hertz).)

- Separation of variables in the Laplacian Δ: To solve $\Delta u = 0$ on a disk of radius a centered at $(0,0)$ in the plane, we begin by writing Δ in polar coordinates and $u(r,\theta) = R(r)\Theta(\theta)$. The equation that results for R, for instance, is Bessel's equation

$$r^2\,R''(r) + r\,R'(r) + (r^2 - k)\,R(r) \;=\; 0$$

where k is a constant.

- Nonlinear mechanics: A pendelum of length ℓ swinging due to the acceleration of gravity g, and making an angle $\theta(t)$ with the verticle at time t, satisfies

$$\ddot{\theta} = -\frac{g}{\ell}\,\sin\theta\,.$$

(This equation neglects air resistance and other losses.) If the motion is restricted to small angles, $\sin\theta \approx \theta$ and the linear equation $\ddot{\theta} = -\frac{g}{\ell}\theta$ is a good approximation.

- Nonlinear mechanics; 2-body equations:

$$\begin{pmatrix}\ddot{x}\\ \ddot{y}\\ \ddot{z}\end{pmatrix} = \frac{-GM}{(x^2+y^2+z^2)^{3/2}}\begin{pmatrix}x\\ y\\ z\end{pmatrix}$$

is the DE for the object with position $(x(t), y(t), z(t))$ orbiting an object of mass M. G is the universal gravitational constant.

- Nonlinear mechanics; n-body equations: Let n particles have positions $\mathbf{x}_i(t) = (x_i(t), y_i(t), z_i(t))$ at time t, $i = 1, 2, ..., n$. Generalizing the gravitational acceleration of the last example to all the bodies we obtain the system of equations

$$\ddot{\mathbf{x}}_i(t) \;=\; -G\sum_{j\neq i}\frac{m_j}{r_{ij}^3}(\mathbf{x}_i(t)-\mathbf{x}_j(t))\,, \qquad i = 1, 2, ..., n\,,$$

where m_j is the mass of the j-th particle and $r_{ij} = |\mathbf{x}_i(t) - \mathbf{x}_j(t)|$.

- The Hodgkin-Huxley equations model the propagation of electric potential (voltage) in nurons and have the form

$$\frac{d}{dt}\begin{pmatrix}v(t)\\ n(t)\\ m(t)\\ h(t)\end{pmatrix} = \begin{pmatrix}C_n n(t)^4(v(t)-K_n) + C_m m(t)^3 h(t)(v(t)-K_m) + C(v(t)-K) + I(t)\\ A_n v(t)(1-n(t)) - B_n v(t) n(t)\\ A_m v(t)(1-m(t)) - B_m v(t) m(t)\\ A_h v(t)(1-h(t)) - B_h v(t) h(t)\end{pmatrix}$$

where v is the voltage, I is the (input) current, n models potassium activation levels, m models sodium activation levels, and h models sodium inactivation levels in the nuron. All other quatities are model constants.

- Linear mechanical and electrical systems without friction or resistence are often modeled by systems of equations of the form

$$M\,\ddot{\mathbf{y}}(t) + K\,\mathbf{y}(t) \;=\; \mathbf{f}(t)$$

where in mechanics $\mathbf{y}$ is a vector of displacements, M is the mass matrix, and K the stiffness matrix, and in electronics $\mathbf{y}$ is a vector of currents, M is an iductance matrix, and K is a (inverse) capacitance matrix. In some applications M and K can change with time; for instance a rocket burns its fuel (mass) as it travels, and a material can become warmer and softer with time (as a device runs).

- The same mechanical and electrical systems as in the preceeding example, but with friction or resistence, are often modeled by equations of the form

$$M\,\ddot{\mathbf{y}}(t) + R\,\dot{\mathbf{y}}(t) + K\,\mathbf{y}(t) \;=\; \mathbf{f}(t)$$

where R is a friction or resistance matrix.

One should be especially aware that the last two examples are ubiquitous for applications in mechanical and electrical engineering.

Turning Higher Order Equations into First Order Systems Any n-th order scalar ODE can be turned into a first order *system* of n equations provided we can solve the equation for the highest order derivative, that is, provided it can be put into the form $y^{(n)}(t) = g(t, y(t), \dot{y}(t), ..., y^{(n-1)}(t))$ for some function g of $n+1$ variables. Let the n-vector valued function $\mathbf{x}(t)$ be

$$\begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ \vdots \\ x_n(t) \end{pmatrix} = \begin{pmatrix} y(t) \\ \dot{y}(t) \\ \ddot{y}(t) \\ \vdots \\ y^{(n-1)}(t) \end{pmatrix} .$$

Then $\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x})$, a first order system, if we set

$$\begin{pmatrix} f_1(t) \\ f_2(t) \\ f_3(t) \\ \vdots \\ f_n(t) \end{pmatrix} = \begin{pmatrix} x_2(t) \\ x_3(t) \\ x_4(t) \\ \vdots \\ g(t, x_1(t), x_2(t), ..., x_n(t)) \end{pmatrix} .$$

We can write any first order, n-dimensional system of ODEs in the form

$$\dot{x} \;=\; f(t, x) \qquad \text{or} \qquad \dot{x}(t) \;=\; f(t, x(t))$$

where $x(t)$ is a $n \times 1$ vector and $f(t, x)$ is an n-vector function of $n+1$ variables.

For a *linear* n-th order equation

$$y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)\dot{y} + a_0(t)y \;=\; h(t)\,,$$

the system of n equations is

$$\dot{\mathbf{x}} = \begin{pmatrix} x_2(t) \\ x_3(t) \\ x_4(t) \\ \vdots \\ -a_{n-1}x_n - a_{n-2}x_{n-1} - \cdots - a_0x_1 + h(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ h(t) \end{pmatrix} .$$

This system, or more general linear systems of ODEs, we write compactly as

$$\dot{x} \;=\; A(t)\,x + f(t)$$

with the understanding that x and f are now (column) vector valued functions of t, and A is an $n \times n$ matrix function of t. If A is a matrix of constants we write $\dot{x} = Ax + f(t)$.

2.1 Linear Constant Coefficient Equations

In this section we will treat first order systems of equations of the form $\dot{x} = Ax + f(t)$ where $x = x(t)$ is $n \times 1$, A is $n \times n$ and constant, and $f(t)$ is a known $n \times 1$ 'forcing' function.

Consider first the 'homogeneous' equation

$$\dot{x}(t) \;=\; Ax(t)\,. \tag{2.1.1}$$

In the case $n = 1$, A is a scalar and the solutions of this equation are known from elementary calculus to be the family of functions $x(t) = e^{At}c$ where c is any constant. When A is a square matrix this same formula can still be used provided we properly interpret the expression e^{At}.[12]

[12] A nice read: Nineteen Dubious Way to Compute the Exponential of a Matrix, *SIAM Review*, vol 20, no 4, Oct 1978.

2.1.1 Theorem (existence of the matrix exponential)**.** *Let A be any $n \times n$ (real or complex) matrix. Then for all $t \in \mathbb{R}$ the $n \times n$ matrix*

$$e^{At} \;=\; \exp(At) \;=\; \sum_{k=0}^{\infty} A^k t^k / k! \tag{2.1.2}$$

is well defined by this absolutely convergent series. Moreover, e^{At} commutes with A and satisfies the matrix differential equation

$$\tfrac{d}{dt} e^{At} = e^{At} A = A e^{At}. \tag{2.1.3}$$

We write e^A for (2.1.2) when $t = 1$.

Proof. For each $m \in \mathbb{N}$ the finite sum $\sum_{k=0}^{m} A^k t^k / k!$ defines an $n \times n$ matrix since each term in the sum is such a matrix. The matrix e^{At} is well defined if we can show that the series (2.1.2) converges in the operator norm of $\mathscr{B}(\mathbb{K}^n, \mathbb{K}^n)$ (Example 1.3.6). It follows from (1.3.7) that

$$\| \sum_{k=0}^{m} A^k t^k / k! \| \leq \sum_{k=0}^{m} \|A\|^k |t|^k / k!$$

for all $m \in \mathbb{N}$. And since $\|A\|$ and $|t|$ are both fixed real numbers the sum on the right is bounded by

$$\sum_{k=0}^{\infty} \|A\|^k |t|^k / k! = e^{\|A\|\,|t|}$$

for all m. The Weierstrass M-test (Proposition 1.1.12) shows that the sequence of finite sums is therefore Cauchy in $\mathscr{B}(\mathbb{K}^n, \mathbb{K}^n)$.

To show A commutes with e^{At} we observe that A commutes with the finite sums

$$A \,(\sum_{k=0}^{m} A^k t^k / k!) = (\sum_{k=0}^{m} A^k t^k / k!)\, A$$

for every m. Now let $m \to \infty$; the limits on both sides must remain equal.

We can differentiate the finite sums term by term:

$$\frac{d}{dt} (\sum_{k=0}^{m} A^k t^k / k!) = \sum_{k=0}^{m} A^k \, k \, t^{k-1} / k! = (\sum_{k=1}^{m} A^{k-1} t^{k-1} / (k-1)!)\, A.$$

(We differentiate a matrix component by component.) By the same argument as above the limit of both sides of this equation exist as $m \to \infty$. Passing to this limit proves (2.1.3). □

2.1.2 Theorem (properties of the matrix exponential)**.** *If A and B are $n \times n$ (real or complex) matrices the following hold.*

(a) $e^A e^B = e^{A+B}$ if and only if $AB = BA$, in which case $e^A e^B = e^B e^A$.

(b) For $t = 0$, $e^{A(0)} = e^0 = I$ where I is the identity.

(c) For any $t \in \mathbb{R}$, $(e^{At})^{-1} = e^{-At} = e^{A(-t)}$.

Proof. $AB = BA$ if and only if the binomial formula holds:

$$(A + B)^m = \sum_{k=0}^{m} \frac{m!}{k!(m-k)!} \, A^k \, B^{m-k}$$

for all $m \in \mathbb{N}$. (The "only if" is trivial; to prove "if" write out the case $m = 2$.) Thus, $AB = BA$ implies

$$\sum_{k=0}^{\infty} \frac{A^k}{k!} \sum_{\ell=0}^{\infty} \frac{B^\ell}{\ell!} = \sum_{m=0}^{\infty} \sum_{k=0}^{m} \frac{1}{k!(m-k)!} \, A^k \, B^{m-k}$$

$$= \sum_{m=0}^{\infty} \frac{1}{m!} \sum_{k=0}^{m} \frac{m!}{k!(m-k)!} \, A^k \, B^{m-k} = \sum_{m=0}^{\infty} \frac{1}{m!} (A + B)^m.$$

Here we have freely changed the order of summation in a doubly infinite series because both series are absolutely convergent for any fixed A and B. The first equality was obtained by changing the double sum over the latice of non-negative integral k, ℓ points in the plane so that we first sum diagonally along lines of the form $k+\ell = m =$ constant. Finally $AB = BA$ if and only if $e^A e^B = e^B e^A$ since both sides equal e^{A+B} exactly when A and B commute.

To complete the proof of (a) we must show that $e^A\, e^B = e^B\, e^A$ implies $AB = BA$. This argument is outside the scope of these notes, but we will sketch it in a footnote.[13]

(b) follows by putting the zero matrix into the power series definition of e^A.

For any $t \in \mathbb{R}$, At obviously commutes with $-At$. Thus, $e^{At}e^{-At} = e^{At-At} = I$ which gives (c). □

2.1.3 Remark. Point (c) of the last theorem shows that e^{At} is non-singular for every square matrix A and $t \in \mathbb{R}$.

2.1.4 Theorem (linear constant coefficient equations)**.** *Let A be an $n \times n$ matrix and $x_0 \in \mathbb{K}^n$. Then the linear constant coefficient initial value problem: find $x \in C^1(\mathbb{R}; \mathbb{K}^n)$ which satisfies*

$$\dot{x}(t) \;=\; Ax(t) \qquad\qquad \textit{and} \qquad\qquad x(0) \;=\; x_0 \tag{2.1.4}$$

has a unique solution

$$x(t) \;=\; e^{At}x_0 \tag{2.1.5}$$

for all $t \in \mathbb{R}$. If, in addition, $f \in C(\mathbb{R}; \mathbb{K}^n)$ then the linear constant coefficient initial value problem: find $x \in C^1(\mathbb{R}; \mathbb{K}^n)$ which satisfies

$$\dot{x}(t) \;=\; Ax(t) + f(t) \qquad\qquad \textit{and} \qquad\qquad x(0) \;=\; x_0 \tag{2.1.6}$$

has a unique solution

$$x(t) = e^{At}x_0 + \int_0^t e^{A(t-s)} f(s)\, ds \tag{2.1.7}$$

for all $t \in \mathbb{R}$.

Proof. The solution (2.1.5) is a special case of (2.1.7), but also follows directly from Theorems 2.1.1 and 2.1.2. To show (2.1.7) we write our ODE as $\dot{x} - Ax = f$ and use an integrating factor to solve the equation. This only works for systems if A is constant. Multiply on the left by the matrix e^{-At} and we have (from the product rule)

$$\tfrac{d}{dt}[e^{-At}x(t)] = e^{-At}f(t).$$

Integrating both sides gives

$$e^{-At}x(t) = x(0) + \int_0^t e^{-As} f(s)\, ds.$$

Multiply this equation on the left by e^{At} to get

$$x(t) = e^{At}x(0) + e^{At}\int_0^t e^{-As} f(s)\, ds$$

which is (2.1.4). □

[13] For any $A \in \mathscr{B}(X)$ and function f which is holomorphic in a neighborhood of the spectrum of A, we can define $f(A)$ using the Riesz functional calculus. (As a first cut we just use the power series definition of f. This may need to be extended by analytic continuation if it does not converge on all of $\sigma(A) \subset \mathbb{C}$.) Now e^A is non-singular, i.e., $0 \notin \sigma(A)$, for all A, and the natural logrithm, $\log(\cdot)$ is holomorphic and singlevalued on any simply connected region not containing 0. So setting $f(\cdot) = log(\cdot)$ means that $log(e^A) = A$ for all $A \in \mathscr{B}(X)$. Further, if $CD = DC$ in $\mathscr{B}(X)$ then $f(C)f(D) = f(D)f(C)$ (just multiply out the two power series). Thus, $\log(e^A)\log(e^B) = \log(e^B)\log(e^A)$, and this is just $AB = BA$.

Jordan Canonical Form If A is a square matrix the matrix e^{At} can be computed explicitly if the Jordan canonical form of A is known. We recall this decomposition of a square matrix.

2.1.5 Definition. A square matrix B is a *Jordan block* if it has the form

$$B = \begin{pmatrix} \lambda & 1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & 0 & \cdots & 0 \\ 0 & 0 & \lambda & 1 & \cdots & 0 \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \lambda & 1 \\ 0 & \cdots & 0 & 0 & 0 & \lambda \end{pmatrix} . \tag{2.1.8}$$

We sometimes write $B = B(\lambda)$ to display the dependence of B on λ. A square matrix J is in *Jordan form* or *Jordan canonical form* if it has the block diagonal form

$$J = \begin{pmatrix} B_1 & 0 & 0 & \cdots & 0 \\ 0 & B_2 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & 0 & B_{\ell-1} & 0 \\ 0 & \cdots & 0 & 0 & B_\ell \end{pmatrix} . \tag{2.1.9}$$

where each B_j is a Jordan block. If $B_j = B(\lambda_j)$ the λ_j's need not be distinct.

2.1.6 Theorem. *If A is an $n \times n$ (real or complex) matrix there is a non-singular $n \times n$ matrix P and a Jordan matrix J such that*

$$A = PJP^{-1} .$$

This decomposition is unique except that the Jordan blocks could appear in any order down the block diagonal of J.

We refer to a linear algebra text for the (non-trivial) proof of this important theorem.
Now we begin the process of computing e^{At}.

2.1.7 Lemma. *If B is a $k \times k$ Jordan block, then*

$$B = \lambda I_k + U_k$$

where I_k is the $k \times k$ identity and U_k is the $k \times k$ matrix with 1's on the first super diagonal and zeros elsewhere. Further, $I_k U_k = U_k I_k$.

Proof. The decomposition of B into the sum of I_k and U_k is obvious. We leave as a simple exercise the verification that I_k and U_k commute. □

2.1.8 Lemma. *Let B, λ, I_k, and U_k be as in the preceding lemma. Then*

$$e^{\lambda I_k t} = \begin{pmatrix} e^{\lambda t} & 0 & \cdots & 0 \\ 0 & e^{\lambda t} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & e^{\lambda t} \end{pmatrix}_{k \times k} = e^{\lambda t} I_k , \tag{2.1.10}$$

$$e^{U_k t} = \begin{pmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{k-2}}{(k-2)!} & \frac{t^{k-1}}{(k-1)!} \\ 0 & 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{k-2}}{(k-2)!} \\ 0 & 0 & 1 & t & \cdots & \frac{t^{k-3}}{(k-3)!} \\ \vdots & & & \ddots & \ddots & \vdots \\ 0 & & \cdots & 0 & 1 & t \\ 0 & & \cdots & 0 & 0 & 1 \end{pmatrix}_{k \times k} , \tag{2.1.11}$$

and

$$e^{Bt} \;=\; e^{\lambda t} e^{U_k t} \;. \tag{2.1.12}$$

For instance

$$e^{U_2 t} \;=\; \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad e^{U_3 t} \;=\; \begin{pmatrix} 1 & t & t^2/2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{pmatrix} \;,$$

and if $B = \lambda I_3 + U_3$ then

$$e^{Bt} \;=\; \begin{pmatrix} e^{\lambda t} & te^{\lambda t} & t^2 e^{\lambda t}/2 \\ 0 & e^{\lambda t} & te^{\lambda t} \\ 0 & 0 & e^{\lambda t} \end{pmatrix} \;.$$

Proof. Equation (2.1.10) is easily verified by substituting $A = \lambda I_k$ into the infinite series for e^{At}. Equation (2.1.11) is also verified by substituting $A = U_k$ into the infinite series for e^{At}. In this case the series is zero after a finite number of terms since U_k is nilpotent. Since λI_k and U_k commute, (2.1.12) follows from Theorem 2.1.2a. □

2.1.9 Example. Let

$$A \;=\; \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix} \qquad \text{and} \qquad f(t) \;=\; \begin{pmatrix} e^{-t} \\ 0 \end{pmatrix} \;.$$

Let's calculate the general solution of the equation

$$\dot{x}(t) \;=\; Ax(t) + f(t) \;.$$

As in the preceding examples we can compute

$$e^{At} \;=\; \begin{pmatrix} e^{-t} & te^{-t} \\ 0 & e^{-t} \end{pmatrix} \;.$$

A 'particular solution' (see Theorem 2.1.4) is

$$\begin{aligned} y(t) \;&=\; e^{At} \int_0^t e^{-As} f(s)\, ds \;=\; e^{At} \int_0^t \begin{pmatrix} e^s & -se^s \\ 0 & e^s \end{pmatrix} \begin{pmatrix} e^{-s} \\ 0 \end{pmatrix} ds \\ &= e^{At} \int_0^t \begin{pmatrix} 1 \\ 0 \end{pmatrix} ds \;=\; e^{At} \begin{pmatrix} t \\ 0 \end{pmatrix} \;=\; \begin{pmatrix} te^{-t} \\ 0 \end{pmatrix} \;. \end{aligned}$$

And the general solution can be written

$$x(t) \;=\; y(t) + e^{At} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \;=\; \begin{pmatrix} te^{-t} \\ 0 \end{pmatrix} + \begin{pmatrix} c_1 e^{-t} + c_2 te^{-t} \\ c_2 e^{-t} \end{pmatrix}$$

for constants c_1 and c_2.

The following result gives us the final part of the formula for e^{At} when we already know the Jordan form of A.

2.1.10 Proposition. *Let A be an $n \times n$ (real or complex) matrix with Jordan form $A = PJP^{-1}$ where J has the form (2.1.9). Then e^{At} has the form*

$$e^{At} \;=\; P \begin{pmatrix} e^{B_1 t} & 0 & 0 & \cdots & 0 \\ 0 & e^{B_2 t} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & 0 & e^{B_{\ell-1} t} & 0 \\ 0 & \cdots & 0 & 0 & e^{B_\ell t} \end{pmatrix} P^{-1} \tag{2.1.13}$$

where each $e^{B_j t}$ has the form (2.1.12).

Proof. One first checks that $A^k = PJ^kP^{-1}$ for any $k \in \mathbb{N}_0$. Then, since scalars can be moved around at will and the matrix multiplications involve only finite sums, we have

$$\sum_0^\infty A^k t^k/k! \;=\; \sum_0^\infty P\,(J^k t^k/k!)\,P^{-1} \;=\; P\sum_0^\infty (J^k t^k/k!)\,P^{-1}\;.$$

Of course, J^k is the block diagonal matrix whose diagonal blocks are B_j^k. □

2.1.11 Exercise. Let $k > 0$. Find the general solution of $((\frac{d}{dt})^2 + k^2)x(t) = 0$ in two ways. (1) By converting this equation to a 2×2 system and computing e^{At}. And (2) from elementary methods, i.e., substitue $x = e^{rt}$ into the equation and solve for r. Compare both answers.

2.1.12 Exercise. Find the general solution of $(\frac{d}{dt} + c)^2 x(t) = 0$ in two ways. (1) By converting this equation to a 2×2 system and computing e^{At}. And (2) from elementary methods, i.e., substitue $x = e^{rt}$ and te^{rt} (reduction of order). Compare both answers.

2.1.13 Exercise. Solve equation (2.1.4) when:

(a)
$$A = \begin{pmatrix} 4 & -3 \\ 8 & -6 \end{pmatrix} \qquad \text{and} \qquad x_0 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}\;,$$

(b)
$$A = \begin{pmatrix} 3 & 2 & 4 \\ 2 & 0 & 2 \\ 4 & 2 & 3 \end{pmatrix} \qquad \text{and} \qquad x_0 = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}\;,$$

(c)
$$A = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 2 & 2 \\ -1 & 1 & 3 \end{pmatrix} \qquad \text{and} \qquad x_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\;,$$

(d)
$$A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & -1 \\ -3 & 2 & 4 \end{pmatrix} = PJP^{-1}$$

where
$$P = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ -1 & 1 & 1 \end{pmatrix}, \qquad J = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix} \qquad \text{and} \qquad x_0 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}\;.$$

2.1.14 Exercise. Show that if the initial condition $x(0) = x_0$ is replaced by $x(t_0) = x_0$ for any $t_0 \in \mathbb{R}$, then the solution is

$$x(t) = e^{A(t-t_0)}x_0 + \int_{t_0}^t e^{A(t-s)} f(s)\,ds \qquad\qquad t \in \mathbb{R}.$$

2.1.15 Exercise. Show that Theorem 2.1.4 remains true if f is only assumed to satisfy $|\int_0^t |f(s)|\,ds| < \infty$ for every $t \in \mathbb{R}$, except that the solution $x(\cdot)$ may no longer be continuously differentiable.

Second Order Systems In mechanical engineering, systems of equations of the form

$$M\ddot{x} + Kx = f \tag{2.1.14}$$

are common. Here M and K are non-singular $n \times n$ matrices (called the mass and stiffness matrices), and $x = x(t)$ and $f = f(t)$ are column n-vectors. We will consider the case when $M = I$ the identity and $K = A^2$ has a 'square root' A. (In principle we can take $M = I$ since (2.1.14) could be multiplied on the left by M^{-1}.)

By analogy with the matrix exponential which solves $\dot{x} - Ax = 0$ we will look for solutions of $\ddot{x} + Kx = 0$ in the form of matrix valued functions $\sin(At)$ and $\cos(At)$ defined by

$$\sin(At) \;=\; \sum_{k=0}^{\infty} (-1)^k \, \frac{A^{2k+1}\, t^{2k+1}}{(2k+1)!} \tag{2.1.15}$$

$$\cos(At) \;=\; \sum_{k=0}^{\infty} (-1)^k \, \frac{A^{2k}\, t^{2k}}{(2k)!} \;. \tag{2.1.16}$$

2.1.16 Exercise. Show that the series defining $\sin(At)$ and $\cos(At)$ are absolutely convergent for every fixed $n \times n$ matrix A and $t \in \mathbb{R}$.

2.1.17 Exercise. Suppose $f(z) = \sum_0^\infty a_k \, z^k$ and $g(z) = \sum_0^\infty b_k \, z^k$ are entire analytic functions. Show that $f(A)$ and $g(A)$ are well defined elements of $\mathscr{B}(\mathbb{C}^n; \mathbb{C}^n)$ whenever A is an $n \times n$ complex matrix.

Show that $f(A) + g(A) = (f+g)(A)$, where $(f+g)(z) = \sum_0^\infty (a_k + b_k) \, z^k$, and that $f(A)g(A) = (fg)(A)$, where $(fg)(z) = \sum_0^\infty c_k \, z^k$ with $c_k = \sum_{j=0}^k a_j \, b_{k-j}$. Conclude that $f(A)\, g(A) = g(A)\, f(A)$.

2.1.18 Exercise. Fix any $t \in \mathbb{R}$, let A be any $n \times n$ complex matrix, and define the $2n \times 2n$ linear transformation

$$\mathbb{A} \;=\; \begin{pmatrix} \sin(At) & \cos(At) \\ \cos(At) & -\sin(At) \end{pmatrix} \;.$$

Use Exercise 2.1.17 to show that $\mathbb{A}^{-1} = \mathbb{A}$.

2.1.19 Exercise. Show that A commutes with both $\sin(At)$ and $\cos(At)$, and that the series defining $\sin(At)$ and $\cos(At)$ may be differentiated term by term. Show that

$$\tfrac{d}{dt} \sin(At) \;=\; A \cos(At) \;=\; \cos(At)\, A$$

and that

$$\tfrac{d}{dt} \cos(At) \;=\; -A \sin(At) \;=\; -\sin(At)\, A.$$

Conclude, as well, that both $\sin(At)$ and $\cos(At)$ satisfy the matrix differential equation

$$\ddot{X}(t) + A^2 X(t) \;=\; 0$$

where $X(t)$ is $n \times n$.

2.1.20 Exercise. Transform the equation (2.1.14) into a $2n \times 2n$ system of the form

$$\frac{d}{dt} \begin{pmatrix} x \\ \dot{x} \end{pmatrix} \;=\; \begin{pmatrix} 0 & I \\ -A^2 & 0 \end{pmatrix} \begin{pmatrix} x \\ \dot{x} \end{pmatrix} \;.$$

Show that this system has a unique solution when give appropriate initial conditions. Can you say anything about the structure of this solution? Can you compute the matrix exponential?

2.1.21 Exercise. Show that the function

$$x(t) \;=\; \cos(At)x_0 + \sin(At)A^{-1}x_1$$

solves equation (2.1.14) when $f(t) = 0$, and has initial conditions $x(0) = x_0$ and $\dot{x}(0) = x_1$.

2.1.22 Exercise. Can you obtain an explicit formula for $\cos(At)$ and $\sin(At)$ from the Jordan form of A, as we did with the matrix exponential function?

2.2 Linear Equations with Analytic Coefficients

We study here the initial value problem (2.4.17)[14]:

$$\dot{x}(t) = A(t)x(t) + f(t) \qquad \text{and} \qquad x(t_0) = x_0 \tag{2.2.1}$$

when the functions $A(t)$ and $f(t)$ are analytic in a neighborhood about t_0. We will demonstrate a solution $x(t)$ of (2.2.1) in terms of a power series expansion

$$x(t) \;=\; \sum_{k=0}^{\infty} c_k\,(t-t_0)^k \tag{2.2.2}$$

where the $n \times 1$ coefficient vectors c_k can be solved recursively.

2.2.1 Theorem. *Assume all components of the arrays $A(t)$ and $f(t)$ are analytic functions at $t_0 \in \mathbb{C}$ whose power series about t_0 have radius of convergence $r > 0$. Then for any $x_0 \in \mathbb{C}^n$ the initial value problem (2.2.1) has an analytic solution (2.2.2) where the power series converges on $\{t \in \mathbb{C}\ ;\ |t - t_0| < r\}$.*

Proof. There is no loss in generality if we assume $t_0 = 0$. For if not we can set $s = t - t_0$; then $d/dt = d/ds$ and any poiwer series of the form (2.2.2) becomes $\sum_{k=0}^{\infty} c_k\, s^k$ with the same coefficients c_k as before.

So assume

$$A(t) \;=\; \sum_{k=0}^{\infty} A_k\, t^k \qquad \text{and} \qquad f(t) \;=\; \sum_{k=0}^{\infty} f_k\, t^k \tag{2.2.3}$$

where each series converges for all $|t| < r$ in the complex plane. Here, each A_k is an $n \times n$ (constant) matrix and each f_k an $n \times 1$ (constant) vector. In order that the initial condition in (2.2.1) be satisfied we set $c_0 = x_0$ in (2.2.2). Now if we substitute (2.2.2) into the differential equation (2.2.1) we can collect like powers of t (taking $t_0 = 0$) to obtain

$$(k+1)\, c_{k+1} \;=\; \big(\sum_{\ell=0}^{k} A_\ell\, c_{k-\ell}\big) + f_k \tag{2.2.4}$$

for all $k \in \mathbb{N}_0$. With c_0 known this equation can be used to recursively solve for all other c_k's.

We must now show that the series (2.2.2) has radius of convergence at least r. We begin by recalling Cauchy's inequality: *If h is analytic in an open set $U \subset \mathbb{C}$ and the closed disk $\{z\ ;\ |z - z_0| \le \rho\}$ is contained in U, then there is an $R > 0$ such that*

$$|h^{(k)}(z_0)| \;\le\; k!R/\rho^k$$

for every $k \in \mathbb{N}_0$. Here, we use the notation $h^{(k)}$ to denote the k-th derivative of h.

Now take any positive number $\rho < r$. Since the coefficients of t^k in the power series expansions for A and f are $A^{(k)}/k!$ and $f^{(k)}/k!$, we conclude that there are positive constants C_A and C_f such that

$$\|A_k\| \;\le\; C_A/\rho^k\ , \qquad\qquad |f_k| \;\le\; C_f/\rho^k \tag{2.2.5}$$

for all $k \in \mathbb{N}_0$. These combined with (2.2.4) show that

$$(k+1)\,|c_{k+1}| \;\le\; \big(\sum_{\ell=0}^{k} \|A_\ell\|\,|c_{k-\ell}|\,\big) + |f_k| \;\le\; \big(\sum_{\ell=0}^{k} \frac{C_A}{\rho^k}\,|c_{k-\ell}|\,\big) + \frac{C_f}{\rho^k} \tag{2.2.6}$$

for all $k \in \mathbb{N}_0$.

Consider a real valued series $\{\xi_k\}_0^\infty$ which satisfies the difference equation

$$(k+1)\xi_{k+1} \;=\; \big(\sum_{\ell=0}^{k} \frac{C_A}{\rho^\ell}\,\xi_{k-\ell}\,\big) + \frac{C_f}{\rho^k} \tag{2.2.7}$$

[14] In this section we follow Henrici, *Applied and Computational Complex Analysis*, volume 2, chapter 9. This reference can be consulted for more extensive information.

for all $k \in \mathbb{N}_0$ together with the initial condition $\xi_0 \geq |c_0|$. Since $\xi_0 > 0$ it is apparent from (2.2.6) that $\xi_k > 0$ for all $k \in \mathbb{N}_0$. We also have

$$|c_1| \;\leq\; \frac{\alpha}{\rho}\,|c_0| + \frac{\beta}{\rho} \;\leq\; \frac{\alpha}{\rho}\,\xi_0 + \frac{\beta}{\rho} \;=\; \xi_1$$

and

$$2\,|c_2| \;\leq\; \frac{\alpha}{\rho}\,\big(\,|c_1| + \frac{|c_0|}{\rho}\,\big) + \frac{\beta}{\rho} \;\leq\; \frac{\alpha}{\rho}\,\big(\,\xi_1 + \frac{\xi_0}{\rho}\,\big) + \frac{\beta}{\rho} \;=\; 2\,\xi_2\ .$$

And in general, by combining (2.2.6) and (2.2.7), we see that

$$|c_k| \;\leq\; \xi_k \tag{2.2.8}$$

for all $k \in \mathbb{N}_0$. If we can determine a radius of convergence for the series

$$w(t) \;=\; \sum_0^\infty \xi_k\, t^k \tag{2.2.9}$$

(2.2.8) shows that (2.2.2) will have at least the same radius of convergence.

The function $w(t)$, if the series (2.2.9) defines one at all, has derivative

$$\dot w \;=\; \sum_{k=0}^\infty (k+1)\,\xi_{k+1}\, t^k \tag{2.2.10}$$

with the same radius of convergence as w (as the ratio test shows). Inserting (2.2.7) into (2.2.10) yields

$$\begin{aligned}\dot w(t) \;&=\; C_A \sum_{k=0}^\infty \sum_{\ell=0}^k \frac{\xi_{k-\ell}\, t^k}{\rho^\ell} \;+\; C_f \sum_{k=0}^\infty \frac{t^k}{\rho^k} \;=\; C_A \sum_{\ell=0}^\infty \Big(\sum_{k=\ell}^\infty \xi_{k-\ell}\, t^{k-\ell}\Big) \frac{t^\ell}{\rho^\ell} + C_f\,\frac{\rho}{\rho - t} \\ &=\; C_A\,\frac{\rho}{\rho - t}\, w(t) + C_f\,\frac{\rho}{\rho - t}\ .\end{aligned} \tag{2.2.11}$$

This is a first order linear equation with integrating factor $(t-\rho)^{\rho\, C_A}$ and solution

$$w(t) \;=\; \big(\, w(0) + \frac{C_f}{C_A}\,\big)\,\big(\,\frac{\rho}{\rho - t}\,\big)^{\rho\, C_A} \;-\; \frac{C_f}{C_A}\ .$$

Thus, $w(t)$ is analytic for $|t| < \rho$ and the series (2.2.9) necessarily has radius of convergence at least ρ. Since $\rho < r$ was arbitrary we conclude that (2.2.2) converges for all $|t| < r$. □

For linear constant coefficient equations the matrix function e^{At} played a special role in that it could be used to wirte down a particular solution of $\dot x = Ax + f(t)$ as $x(t) = \int_0^t e^{A(t-s)}\, f(s)\, ds$. In this context, such a matrix valued function is called a fundumental solution. Fundumental solutions exist for equations with analytic coefficients also. This is explored in the following

2.2.2 Exercise. Let $W(t) = \sum_0^\infty W_k\, t^k$ be an $n \times n$ matrix-valued function which solves the matrix differential equation $\dot W \;=\; A(t)\, W$. Obtain the recurrence relation for the coefficients W_k if $A(t) = \sum_0^\infty A_k t^k$ is analytic near $t = 0$. Use the initial condition $W(0) = I$. Show that the function $x(t) = \int_0^t W(t-s)\, f(s)$ solves the equation $\dot x = A(t)x + f(t)$ whether or not $f(t)$ is analytic.

Systems with Singular Points In many applications to the boundary value problems of mathematical physics one wants to solve equations like (2.2.1) where the entries of A are analytic in a neighborhood of t_0, but have a singularity at the point t_0 itself. For instance, the method of separating variables for Laplacian's equation, $\Delta u = 0$, in cylindrical coordinates, where $u(r, \theta, z) = R(r)\, H(\theta)\, Z(z)$, leads to an ordinary differential equation of the form

$$r^2\, R'' + r\, R' + (\lambda r^2 - \nu^2)\, R \;=\; 0$$

where $' = d/dr$. This is Bessel's equation with parameters λ and ν. If we set $S(r) = R'(r)$ we can write this equation as the first order system

$$\begin{pmatrix} R \\ S \end{pmatrix}' \;=\; \begin{pmatrix} 0 & 1 \\ \frac{\nu^2}{r^2} - \lambda & -\frac{1}{r} \end{pmatrix} \begin{pmatrix} R \\ S \end{pmatrix} .$$

Bessel's equation does not have analytic coefficients in a neighborhood of $r = 0$, where a solution is usually sought. We will state a theorem that applies to many of these problems. The structure of the fundamental matrix must be expressed in terms of *complex* values of t.

Let's begin with a simple example that illustrates the theory. Suppose $a > 0$ is a constant. The equation

$$\dot{x}(t) - \frac{a}{t}\, x(t) \;=\; 0$$

has an integrating factor t^{-a} which turns it into

$$\tfrac{d}{dt}(t^{-a}\, x) \;=\; 0$$

with solution $x(t) = c\, t^a$. Here c is the constant of integration. If $t = z$ is a complex variable one defines z^a by $e^{a\, \log(z)}$ where $\log(z) = \log(r) + i\theta$ when we use the notation $z = re^{i\theta}$. (The expression $\log(r)$ is the familiar logarithm of a non-negative real number.)

Because θ may be replaced by $\theta + 2\pi k$ for any $k \in \mathbb{Z}$ in the expression $z = re^{i\theta}$, there is a $2\pi k$ ambiguity in the value of $\log(z)$ unless we fix a 'branch' of this function by taking a 'cut' along any ray from 0 to infinity in the complex plane. $\log(z)$ is then a single valued analytic function if we fix its value (the angle θ) at some point z_0 not on the cut, so long as we do not cross the cut as z varies. If the constant a is an integer the function $z^a = e^{a\, \log(z)}$ is not multivalued because e^z is periodic with period $2\pi i$; no branch cut need be made for z^a to be well defined. But in general z^a with also be multivalued. (Example: $z^{1/2} = \sqrt{z}$.)

The form of the fundumental solutions of analytic systems with singularities at $0 \in \mathbb{C}$ is given by

2.2.3 Theorem. *Let $A(t)$ be analytic in the punctured disk $D = \{t \in \mathbb{C} \;;\; 0 < |t| < r\}$. Then there is a, possibly multivalued, fundamental matrix solution $W(t)$ of*

$$\frac{d}{dt} W(t) \;=\; A(t) W(t)$$

at each point t in D. (Such a W already exists locally at each $t \in D$ by Theorem 2.2.1.) Furthermore, there is an $n \times n$ (constant) matrix P and an $n \times n$ matrix valued function $Q(t)$, analytic in D, such that

$$W(t) \;=\; Q(t)\, t^P$$

for all $t \in D$. Here we define

$$t^P = e^{P\, \log(t)} \tag{2.2.12}$$

for $t \in D$, and (the possibly multivalued function) $e^{P\, \log(t)}$ by the matrix power series for the matrix exponential.

Thus the multivalued character of $W(t)$ is a reflection of the multivalued quality of $\log(t)$ in (2.2.12).

Proof. We refer the reader to Henrici, Theorems 9.4a and 9.4b, for the proof. □

2.2.4 Example (Legendre equation)**.** The Legendre equation with parameter $n \in \mathbb{N}_0$ is

$$(1 - x^2)y'' - 2xy' + n(n+1)y \;=\; 0\,.$$

We may write this equation as a 2×2 system with $z = y'$:

$$\begin{pmatrix} y \\ z \end{pmatrix}' \;=\; \begin{pmatrix} 0 & 1 \\ \frac{n(n+1)}{1-x^2} & -\frac{2x}{1-x^2} \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} .$$

The point $x_0 = 0$ is an ordinary point, at which the matrix $A(x)$ is analytic, so we may solve this equation with a power series FINISH

2.2.5 Example (Bessel's equation)**.** Bessel's equation with parameter ν is

$$x^2y'' + xy' + (x^2 - \nu^2)y \;=\; 0\,.$$

Writing this as a second order equation gives

$$\begin{pmatrix} y \\ z \end{pmatrix}' \;=\; \begin{pmatrix} 0 & 1 \\ \frac{\nu^2 - x^2}{x^2} & -\frac{1}{x} \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix}$$

where $z = y'$. We wish to solve this equation with a series expansion about $x_0 = 0$. FINISH COMPUTATIONS

2.2.6 Exercise. Consider the equation $y'' + y = 0$ whose solutions are $\sin x$ and $\cos x$. Write this equations as a 2×2 system and solve using vector valued power series.

2.2.7 Exercise. Consider Airy's equation $y'' - xy = 0$. Write this equations as a 2×2 system and solve using vector valued power series.

2.2.8 Exercise. Solve the system

$$\begin{pmatrix} y \\ z \end{pmatrix}' \;=\; \begin{pmatrix} 0 & 1 \\ -x & -b \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix}$$

with a power series about the point $x_0 = 0$.

2.2.9 Exercise. Solve the system

$$\begin{pmatrix} y \\ z \end{pmatrix}' \;=\; \begin{pmatrix} 1 & x \\ -x^2 & 1 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix}$$

with a power series about the point $x_0 = 0$.

2.3 Laplace Transform Methods

Let $x(t)$ be an $\mathbb{R}^n$ valued function of t and define, when the integral exists, the Laplace transform

$$\mathscr{L}x(s) \;=\; X(s) \;=\; \int_0^\infty e^{-st}\, x(t)\, dt\,.$$

The integral is of course applied to each component of $e^{-st}x(t)$. It is linear in the sense that if A and B are $m \times n$ matrices and $x(t)$ and $y(t)$ are $\mathbb{R}^n$ valued functions, then

$$\mathscr{L}(Ax + By)(s) \;=\; AX(s) + BY(s)\,.$$

If $x(t)$ is complex valued the transform may be applied to the real and imaginary parts.

2.3.1 Example. Here are a few elementary Laplace transforms of scalar functions when $a, b \in \mathbb{R}$ and $n \in \mathbb{N}$.

- $\mathscr{L}(1) \;=\; \frac{1}{s}$
- $\mathscr{L}(e^{at}) \;=\; \frac{1}{s-a}$
- $\mathscr{L}(t^n) \;=\; \frac{n!}{s^{s+1}}$
- $\mathscr{L}(\sin\, at) \;=\; \frac{a}{s^2+a^2}$
- $\mathscr{L}(\cos\, at) \;=\; \frac{s}{s^2+a^2}$
- $\mathscr{L}(e^{at} \sin\, bt) \;=\; \frac{b}{(s-a)^2+b^2}$
- $\mathscr{L}(e^{at} \cos\, bt) \;=\; \frac{s-a}{(s-a)^2+b^2}$

2.3.2 Theorem (Existence of Laplace Transform)**.** *If each component of $x(t)$ satisfies*

$$|x_j(t)| \;\le\; C\,e^{rt}$$

for some $C > 0$ and $r \in \mathbb{R}$, then the Laplace transform of $x(t)$ exists for all $s \in \mathbb{C}$ such that $\Re s > r$. Moreover $X(s)$ is analytic on this subset of $\mathbb{C}$.

Proof. We leave this proof as an exercise. To show that $X(s)$ is analytic, differentiate under the integral sign to verify the Cauchy-Riemann equations. □

2.3.3 Theorem (Uniqueness of Laplace Transform)**.** *Let $x(t)$ and $y(t)$ be continuous scalar-valued functions defined on $0 \le t < \infty$, and assume $\mathscr{L}(x)$ and $\mathscr{L}(y)$ exist and are equal on a set of the form $\{\Re s > r\}$ in $\mathbb{C}$ for some $r \in \mathbb{R}$. Then $x(t) = y(t)$ for all $t \in [0, \infty)$.*

Proof. This proof is beyond the scope of these notes. But see Marsden, *Basic Complex Analysis*, Section 8.1 for the statement of this theorem (Theorem 2) and a nice discussion. □

2.3.4 Theorem (Laplace Transform of Derivative)**.** *If the Laplace transform of $x(t)$ and $\dot{x}(t)$ exist on $\{s \in \mathbb{C}\ ;\ \Re s > a\}$ then*

$$\mathscr{L}(\dot{x})(s) \;=\; s\,X(s) - x(0)\,.$$

Proof. Integrate by parts:

$$\int_0^\infty e^{-st}\,\dot{x}(t)\,dt = [e^{-st}x(t)]_0^\infty - \int_0^\infty -se^{-st}\,x(t)\,dt = -x(0) + sX(s)$$

where we have infered that $e^{-st}x(t) \to 0$ as $t \to \infty$ since the transform exists when $\Re s$ is sufficiently large. □

One of the most important things about the Laplace transform is the elegant way it handles step and impulse functions. These functions represent forces that are switched on and off, as well as sudden forces, in mechanical and electrical systems. The following theorem will give some valuable formulas related to these functions.

2.3.5 Theorem. *Let $H(t) = 1_{[0,\infty)}(t)$ be the unit (Heaviside) step function and assume $\mathscr{L}(x) = X(s)$ exists for $s > a \ge 0$.*
If $c > 0$ then $\mathscr{L}(H(t-c)x(t-c)) = e^{-cs}X(s)$.
If $c \in \mathbb{R}$ then $\mathscr{L}(e^{ct}x(t)) = X(s-c)$ when $s > a + c$.
If $\tau > 0$ then $\mathscr{L}(\delta(t-\tau)) = e^{-\tau s}$.

Proof. □

Now we consider using the Laplace transform to solve linear constant coefficient equations.

2.3.6 Theorem. *Let $A \in \mathbb{R}^{n\times n}$ and $f(t)$ be an $\mathbb{R}^n$-valued function having a Laplace transform for $s > a$ where $a \in \mathbb{R}$. Then the initial value problem*

$$\dot{x} \;=\; Ax + f(t)\,, \qquad x(0) \;=\; x_0 \tag{2.3.13}$$

has a unique $\mathbb{R}^n$-valued solution given by

$$x(t) \;=\; \mathscr{L}^{-1}(\,(sI - A)^{-1}(F(s) + x_0)\,)\,. \tag{2.3.14}$$

Since $(sI - A)^{-1}$ is a rational matrix function in s, the right side of (2.3.14) can in principle be calculated...
The proof is a simple calculation and an appeal to the uniqueness of the Laplace transform.

Proof. Take the Laplace transform of both sides of (2.3.13) to get

$$sX(s) - x_0 \;=\; AX(s) + F(s)\,.$$

Thus,

$$(sI - A)X(s) \;=\; F(s) + x_0\,.$$

□

2.3.7 Example (Nagle, Saff, & Snider, p. 599 (not used with permission)). Consider the initial value problem

$$\dot{x} = \begin{pmatrix} 0 & 2 \\ -1 & 3 \end{pmatrix} x\,, \qquad x(0) = \begin{pmatrix} -1 \\ 3 \end{pmatrix}\,.$$

Taking the Laplace transform yields the equation

$$\begin{pmatrix} s & -2 \\ 1 & s-3 \end{pmatrix} X(s) \;=\; \begin{pmatrix} -1 \\ 3 \end{pmatrix}\,.$$

Since

$$\begin{pmatrix} s & -2 \\ 1 & s-3 \end{pmatrix}^{-1} \;=\; \frac{1}{(s-1)(s-2)} \begin{pmatrix} s-3 & 2 \\ -1 & s \end{pmatrix}$$

we have

$$X(s) \;=\; \frac{1}{(s-2)(s-1)} \begin{pmatrix} -s+9 \\ 3s+1 \end{pmatrix} \;=\; \begin{pmatrix} \frac{7}{s-2} - \frac{8}{s-1} \\ \frac{7}{s-2} - \frac{4}{s-1} \end{pmatrix}\,.$$

Inverting (consult a table of Laplace transforms or use the examples given above) we have $x(t) = (7e^{2t} - 8e^t\ ,\ 7e^{2t} - 4e^t)'$. The reader may check that the differential equation and the initial conditions are satisfied.

2.3.8 Example (Nagle, Saff, & Snider, p. 599 (not used with permission)). Consider the initial value problem

$$\dot{x} = \begin{pmatrix} 3 & -2 \\ 4 & -1 \end{pmatrix} x + \begin{pmatrix} \sin t \\ -\cos t \end{pmatrix}\,, \qquad x(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}\,.$$

Taking the Laplace transform yields the equation

$$\begin{pmatrix} s-3 & 2 \\ -4 & s+1 \end{pmatrix} X(s) \;=\; \begin{pmatrix} 1/(s^2+1) \\ -s/(s^2+1) \end{pmatrix}\,.$$

Since

$$\begin{pmatrix} s-3 & 2 \\ -4 & s+1 \end{pmatrix}^{-1} \;=\; \frac{1}{(s-1)^2+4} \begin{pmatrix} s+1 & -2 \\ 4 & s-3 \end{pmatrix}$$

we find, after a rather tedious partial fractions expansion, that

$$X(s) \;=\; \frac{1}{s^2+1}\,\frac{1}{(s-1)^2+4} \begin{pmatrix} 3s+1 \\ -(s-4)(s+1) \end{pmatrix} \;=\; \begin{pmatrix} \frac{1}{6}\left(\frac{-7s}{s^2+1} + \frac{5}{s^2+1} + \frac{7s}{(s-1)^2+4} + \frac{-19}{(s-1)^2+4}\right) \\ -\frac{1}{10}\left(\frac{-11s}{s^2+1} + \frac{-7}{s^2+1} + \frac{11s}{(s-1)^2+4} + \frac{-5}{(s-1)^2+4}\right) \end{pmatrix}\,.$$

Inverting (use the examples given above and the second formula in Theorem 2.3.5) we have

$$x(t) \;=\; \begin{pmatrix} (-7\cos t + 5\sin t + 7e^t\cos 2t - 6e^t \sin 2t)/6 \\ (11\cos t + 7\sin t - 11e^t\cos 2t - 3e^t\sin 2t)/10 \end{pmatrix}\,.$$

The reader may check that the differential equation and the initial conditions are satisfied.

2.3.9 Example. Consider the initial value problem

$$\dot{x} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ g(t) \end{pmatrix}\,, \qquad x(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}\,,$$

where $g(t) = H(t-\pi/2) + \delta(t-\pi) - H(t-3\pi/2)$. Using the Laplace transform formulas

$$\mathscr{L}(H(t-\pi/2)) = \frac{e^{-\pi s/2}}{s}, \qquad \mathscr{L}(\delta(t-\pi)) = e^{-\pi s}, \qquad \mathscr{L}(H(t-3\pi/2)) = \frac{e^{-3\pi s/2}}{s},$$

the Laplace transform $X(s)$ satisfies

$$\begin{pmatrix} s & -1 \\ 1 & s \end{pmatrix} X(s) = \begin{pmatrix} 0 \\ G(s) \end{pmatrix}\,, \quad \text{or} \quad X(s) = \begin{pmatrix} \frac{e^{-\pi s/2}}{s(s^2+1)} + \frac{e^{-\pi s}}{s^2+1} - \frac{e^{-3\pi s/2}}{s(s^2+1)} \\ \frac{e^{-\pi s/2}}{s^2+1} + \frac{s\,e^{-\pi s}}{s^2+1} - \frac{e^{-3\pi s/2}}{s^2+1} \end{pmatrix}\,.$$

Using the partial fractions expansion $\frac{1}{s(s^2+1)} = \frac{-s}{s^2+1} + \frac{1}{s}$ we can simplify two terms in the first component of $X(s)$, and invert to obtain

$$x(t) \;=\; \begin{pmatrix} [1-\cos(t-\pi/2)]H(t-\pi/2) + \sin(t-\pi)H(t-\pi) - [1-\cos(t-3\pi/2)]H(t-3\pi/2) \\ \sin(t-\pi/2)H(t-\pi/2) + \cos(t-\pi)H(t-\pi) - \sin(t-3\pi/2)H(t-3\pi/2) \end{pmatrix} .$$

Stability is an issue here (roots inside or outside unit disc), and applications to electric circuits. [15]

2.4 General Existence Theorems and Successive Approximations

We begin with a local existence theorem. For $x \in \mathbb{R}^n$ and $r > 0$ let $B(x,r) = \{y \in \mathbb{R}^n \;;\; |y-x| < r\}$ with closure $\overline{B}(x,r) = \{y \in \mathbb{R}^n \;;\; |y-x| \le r\}$.

2.4.1 Theorem (first local existence)**.** *Let $t_0 \in \mathbb{R}$, $x_0 \in \mathbb{R}^n$, $a > 0$, and $r > 0$; and let*

$$f : [t_0 - a, t_0 + a] \times \overline{B}(x_0, r) \to \mathbb{R}^n$$

be continuous, and Lipschitz continuous in the x variable in $\overline{B}(x_0,r)$: there is a $C > 0$ such that

$$|f(t,x) - f(t,y)| \;\le\; C\,|x-y| \tag{2.4.1}$$

for all $t \in [t_0 - a, t_0 + a]$ and $x, y \in \overline{B}(x_0,r)$. Set $C_0 = \sup\{|f(t,x)| \;;\; (t,x) \in [t_0-a, t_0+a] \times \overline{B}(x_0,r)\}$. Then for any positive $b < \min\{a, r/C_0, 1/C\}$ there is a unique function $x \in C^1([t_0-b, t_0+b]; \mathbb{R}^n)$ such that

$$\dot{x}(t) \;=\; f(t, x(t)) \tag{2.4.2}$$

for all $t \in [t_0 - b, t_0 + b]$ and

$$x(t_0) = x_0 \;. \tag{2.4.3}$$

Furthermore, $x([t_0-b, t_0+b]) \subset \overline{B}(x_0,r)$.

The existence of one sided limits $x((t_0-b)+)$, $x((t_0+b)-)$, $\dot{x}((t_0-b)+)$, and $\dot{x}((t_0+b)-)$ is part of the conclusion of the theorem.

Proof. The solution of (2.4.2) and (2.4.3) will be constructed by the method of *successive approximations*. The form of the equation must first be changed to

$$x(t) \;=\; x_0 + \int_{t_0}^{t} f(s, x(s))\, ds \;. \tag{2.4.4}$$

Equations (2.4.2) and (2.4.3) are equivalent to (2.4.4) by the fundumental theorem of calculus.

A sequence of functions that approximates the solution $x(t)$ is constructed as follows. Let the first approximation be the constant function

$$x_1(t) \;=\; x_0$$

for all $t \in [t_0 - b, t_0 + b]$. The next approximation is

$$x_2(t) \;=\; x_0 + \int_{t_0}^{t} f(s, x_1(s))\, ds$$

for $t \in [t_0 - b, t_0 + b]$. In general the n-th approximation is

$$x_n(t) \;=\; x_0 + \int_{t_0}^{t} f(s, x_{n-1}(s))\, ds$$

[15] TO STEVE: Do feedback systems and the Nyquist stability criterion if possible.

when $n \geq 2$.

Now we must prove that the sequence $x_n(t)$ converges to a solution of (2.4.4). Define

$$F : C([t_0 - b, t_0 + b]; \mathbb{R}^n) \to C([t_0 - b, t_0 + b]; \mathbb{R}^n) \qquad \text{by} \qquad F(z)(t) = x_0 + \int_{t_0}^{t} f(s, z(s))\, ds \ .$$

The solution x of (2.4.4) is a fixed point of F, i.e., satisfies $F(x)(t) = x(t)$ for all $t \in [t_0 - b, t_0 + b]$. And the sequence x_n satisfies $x_n = F(x_{n-1})$. These are the same iterations used in the proof of the contraction mapping theorem.

We will show that F is a contraction on a closed subset of the Banach space $C([t_0 - b, t_0 + b]; \mathbb{R}^n)$. Let

$$Y = \{y \in C([t_0 - b, t_0 + b]; \mathbb{R}^n) \ ; \ y(t_0) = x_0 \text{ and } y(t) \in \overline{B}(x_0, r) \text{ for all } t \in [t_0 - b, t_0 + b]\} \ .$$

If $t \in [t_0 - b, t_0 + b]$ it is easy to see that the evaluation linear functional, $\delta_t : y \mapsto y(t) \in \mathbb{R}^n$, is continuous on Y (see Example 1.3.13). Since $\overline{B}(x_0, r) \subset \mathbb{R}^n$ is closed, $Y_t = \{y \in C([t_0 - b, t_0 + b]; \mathbb{R}^n) \ ; \ y(t) \in \overline{B}(x_0, r)\}$ is closed. Since $Y = \cap_{t \in [t_0 - b, t_0 + b]} Y_t$, Y is also a closed subset of $C([t_0 - b, t_0 + b]; \mathbb{R}^n)$.

Now we claim that $F : Y \to Y$, and that F is a contraction on Y. Let $y \in Y$. Then

$$|F(y)(t) - x_0| \ \leq \ \int_{t_0}^{t} |f(s, y(s))|\, ds \ \leq \ b\, C_0 \ < \ r \ ;$$

this holds for all $t \in [t_0 - b, t_0 + b]$ since $b < r/C_0$, so $F(Y) \subset Y$. Next, let $y, z \in Y$. Then

$$\begin{aligned} |F(y)(t) - F(z)(t)| \ &\leq \ \int_{t_0}^{t} |f(s, y(s)) - f(s, z(s))|\, ds \\ &\leq \ C \int_{t_0}^{t} |y(s) - z(s)|\, ds \ \leq \ C\, b \sup_{s \in [t_0 - b, t_0 + b]} |y(s) - z(s)| \ . \end{aligned}$$

The right side is independent of t so $|F(y) - F(z)|_0 \leq C\, b\, |y - z|_0$ where $|f(t)|_0 = \sup_{t_0 - b \leq t \leq t_0 + b} |f(t)|$ is the norm on $C([t_0 - b, t_0 + b]; \mathbb{R}^n)$. Since $b < 1/C$, F is a contraction on Y and there is a unique solution x of (2.4.4).

Finally, we show that this x is continuously differentiable on $[-b, b]$ and that the end point limits exist. Since x is continuous and satisfies (2.4.4), and hence (2.4.2), $\dot{x}$ is continuous because the right side of (2.4.2) is. The integrand $f(t, x(t))$ in (2.4.4) is continuous, hence bounded, on the compact interval $[t_0 - b, t_0 + b]$, so the left and right hand limits of (2.4.4) exist at $t = t_0 + b$ and $t = t_0 - b$. And since $x(t)$ has left and right limits at $t = t_0 + b$ and $t = t_0 - b$, (2.4.2) shows the same is true of $\dot{x}(t)$. □

In the next theorem we required f to be Lipschitz continuous on all of $\mathbb{R}^n$, not just $\overline{B}(x, r)$; this restriction on f will result in a solution $x(t)$ for (2.4.2) on $-\infty < t < \infty$.

2.4.2 Theorem (second local existence)**.** *Let $t_0 \in \mathbb{R}$ and $a > 0$, and let $f : [t_0 - a, t_0 + a] \times \mathbb{R}^n \to \mathbb{R}^n$ be continuous. Assume there exists a $C_1 > 0$ such that*

$$|f(t, x) - f(t, y)| \ \leq \ C_1\, |x - y| \tag{2.4.5}$$

for all $t \in [t_0 - a, t_0 + a]$ and $x, y \in \mathbb{R}^n$. Then for any positive $b < \min\{a, 1/C_1\}$ and any $x_0 \in \mathbb{R}^n$ there is a unique function $x \in C^1([t_0 - b, t_0 + b]; \mathbb{R}^n)$ such that

$$\dot{x}(t) \ = \ f(t, x(t)) \tag{2.4.6}$$

for all $t \in [t_0 - b, t_0 + b]$, and

$$x(t_0) = x_0 \ . \tag{2.4.7}$$

Proof. The proof is almost the same as the proof of the first local existence theorem. The initial value problem (2.4.6)-(2.4.7) is equivalent to (2.4.4). Define

$$F(z)(t) = x_0 + \int_{t_0}^{t} f(s, z(s))\, ds \ ,$$

and $Y = \{y \in C([t_0 - b, t_0 + b]; \mathbb{R}^n) \; ; \; y(t_0) = x_0\}$. This Y is slightly different than the preceding proof. Y is a closed subset of the Banach space $C([t_0 - b, t_0 + b]; \mathbb{R}^n)$ because it is the inverse image of the closed set $\{x_0\}$ of the (continuous) evaluation functional $\delta_{t_0} : y \mapsto y(t_0)$.

We first show that, for any $b \in (0, a]$, $F : Y \to Y$, i.e., that $F(Y) \subset Y$. Obviously $F(z)(t_0) = x_0$ whatever $z \in Y$. Set

$$C_0 \;=\; \sup\{|f(t, x_0)| \; ; \; t_0 - a \le t \le t_0 + a\} \; .$$

Since $[t_0 - a, t_0 + a]$ is compact and $t \mapsto |f(t, x_0)|$ is continuous such a C_0 exists. For any $t \in [t_0 - a, t_0 + a]$ and $x \in Y$

$$|f(t, x)| \;\le\; |f(t, x_0| + |f(t, x) - f(t, x_0)| \;\le\; |f(t, x_0| + C_1 \, |x - x_0| \; . \tag{2.4.8}$$

We now obtain

$$\begin{aligned} |F(x)(t) - F(x)(s)| \;\le\; \operatorname{sign}(t - s) \int_s^t |f(r, x(r))| \, dr \;\le\; \operatorname{sign}(t - s) \int_s^t |f(r, x_0)| + C_1 \, |x(r) - x_0| \, dr \\ \le\; C_0 \, |t - s| + C_1 \operatorname{sign}(t - s) \int_s^t |x(r) - x_0| \, dr \end{aligned} \tag{2.4.9}$$

where $\operatorname{sign}(t) = 1$ if $t \ge 0$ and $= -1$ if $t < 0$. Since $x \in C([t_0 - b, t_0 + b])$ the right side goes to zero as $s \to t$ for any $t \in [t_0 - b, t_0 + b]$. So $F(x) \in Y$.

Finally, if b is sufficiently small, we show that F is a contraction on Y. We have

$$\begin{aligned} |F(x)(t) - F(y)(t)| \;\le\; \operatorname{sign}(t - t_0) \int_{t_0}^t |f(s, x(s)) - f(s, y(s))| \, ds \;\le\; C_1 \operatorname{sign}(t - t_0) \int_{t_0}^t |x(s) - y(s)| \, ds \\ \le C_1 \left(\Big| \int_{t_0}^t ds \Big| \right) \Big(\sup_{s \in [t_0, t] \text{ or } [t, t_0]} |x(s) - y(s)| \Big) \;\le\; C_1 \, b \, |x - y|_0 \end{aligned}$$

where $| \cdot |_0$ is the norm on $C([t_0 - b, t_0 + b]; \mathbb{R}^n)$. If $b < \min\{a, 1/C_1\}$, F is a contraction. □

2.4.3 Corollary. *Let $t_0 \in \mathbb{R}$ and $a > 0$, and let $f : [t_0, t_0 + a] \times \mathbb{R}^n \to \mathbb{R}^n$ be continuous. Assume there exists a $C_1 > 0$ such that*

$$|f(t, x) - f(t, y)| \;\le\; C_1 \, |x - y|$$

for all $t \in [t_0, t_0 + a]$ and $x, y \in \mathbb{R}^n$. Then for any positive $b < \min\{a, 1/C_1\}$ and any $x_0 \in \mathbb{R}^n$ there is a unique $x \in C^1([t_0, t_0 + b]; \mathbb{R}^n)$ such that

$$\dot{x}(t) \;=\; f(t, x(t))$$

for all $t \in [t_0, t_0 + b]$, and

$$x(t_0) = x_0 \; .$$

Proof. The proof may be constructed almost exactly as the proof of Theorem 2.4.2. But we may also set

$$\tilde{f}(t) \;=\; \begin{cases} f(t) & t \ge 0 \; , \\ f(-t) & t < 0 \; . \end{cases} \tag{2.4.10}$$

This $\tilde{f}$ satisfies the hypothesis of Theorem 2.4.2; we conslude that there is a unique solution $\tilde{x}$ on the interval $t_0 - a \le t \le t_0 + a$. The function $x(t) = \tilde{x}(t)$ for $0 \le t \le a$ now satisfies the corollary. □

If $U \subset \mathbb{R}^m$, a function $f : U \to \mathbb{R}^n$ is said to be *Lipschitz continuous* on U if there is a $C > 0$ such that

$$|f(x) - f(y)| \le C|x - y| \tag{2.4.11}$$

for all x and y in U. The function $f(x) = x^2$ on $-\infty < x < \infty$ is not Lipschitz continuous on $\mathbb{R}$, but it is Lipschitz continuous on every bounded subset of $\mathbb{R}$.

The following result is often useful for checking the hypotheses (2.4.1) and (2.4.5).

2.4.4 Theorem. *Let $U \subset \mathbb{R}^m$ be convex and open, and let $f : U \to \mathbb{R}^n$ be differentiable on U. If there is a $C > 0$ such that $\|Df(x)\| \le C$ for all $x \in U$ (the operator norm of the matrix $[\partial_i f_j]$), then f satisfies (2.4.11) for all $x, y \in U$.*

Proof. If x and y are in U the mean value theorem shows that

$$|f(y) - f(x)| \le |Df((1-t)x + ty)[y - x]|$$

for some $t \in (0, 1)$. We have used the convexity of U to insure the segment connecting x and y lies in U. But we also have

$$|Df((1-t)x + ty)[y - x]| \le \|Df((1-t)x + ty)\| \, |y - x| \le C \, |y - x| \, .$$

□

The ball $B(x, r)$ in $\mathbb{R}^n$ is convex since $|(1-t)x + ty| \le (1-t)|x| + t|y|$ for $0 \le t \le 1$. This fact will be combined with the preceding result in some of the theorems below.

2.4.5 Theorem (global existence)**.** *Let $T > 0$, and $f : [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$ be continuous and satisfy the following Lipschitz continuity condition on $\mathbb{R}^n$: there is a constant $C_1 > 0$ such that*

$$|f(t, x) - f(t, y)| \le C_1 \, |x - y| \tag{2.4.12}$$

holds for all $x, y \in \mathbb{R}^n$ and $t \in [0, T]$. Then, for any $x_0 \in \mathbb{R}^n$ the initial value problem

$$\dot{x}(t) = f(t, x(t)) \tag{2.4.13}$$

for $t \in [0, T]$ and

$$x(0) = x_0 \tag{2.4.14}$$

has a unique solution $x(\cdot) \in C^1([0, T]; \mathbb{R}^n)$.

First proof. If $T < 1/C_1$ then Corollary 2.4.3 holds with $b = T$ and the theorem is proved. If $T \ge 1/C_1$ choose any $b \in (0, 1/C_1)$ and partition the interval $[0, T]$ into m subintervals of length b or shorter. Let this partition be $[0, t_1]$, $[t_1, t_2]$, ..., $[t_{m-1}, t_m]$.

Corollary 2.4.3 shows that (2.4.13) can be solved successively on the subintervals $[t_k, t_{k+1}]$. On $[0, t_1]$ we use the initial value $x(0) = x_0$; then on $[t_1, t_2]$ we use the newly computed value $x(t_1)$ as initial condition. Continuing through all subintervals we obtain a solution $x(t) \in C([0, T]; \mathbb{R}^n)$. Here we use the hypothesis that the same constant C_1 works for all subintervals.

It is clear from the corollary that $\dot{x}(t)$ is continuous on $[0, T]$ except possibly at the end points t_k of the subintervals. To check that $\dot{x}(t)$ is continuous there, let $x_k(t)$ be the solution obtained from the corollary on $[t_{k-1}, t_k]$, and $x_{k+1}(t)$ the one on $[t_k, t_{k+1}]$. For $t < t_k$, $x(t) = x_k(t) = \int_{t_{k-1}}^{t} f(s, x_k(s)) \, ds$ and $\dot{x}(t) = f(t, x_k(t))$; and for $t > t_k$, $\dot{x}(t) = f(t, x_{k+1}(t))$. Since we know that $x(t)$ is continuous at t_k, that is, $x_k(t_k-) = x_{k+1}(t_k+)$, it is also clear that $\dot{x}_k(t_k-) = \dot{x}_{k+1}(t_k+)$. □

We know from the most elementary differential equation, $\dot{x} = ax$ for $a \in \mathbb{R}$, that solutions of ordinary differential equations can grow exponentially. This fact suggests that we may be able to define a contraction mapping that solves the system (2.4.13)-(2.4.14) on the whole interval $[0, T]$ if we use an exponential decaying weight in the Banach space norm for $C([0, T]; \mathbb{R}^n)$. The following proof shows that this is so.

Second proof. For $\rho > 0$, to be specified later, consider the norm

$$|x|_\rho = \sup_{0 \le t \le T} e^{-\rho t} |x(t)|$$

for $x \in C([0, T]; X)$. The norm $|\cdot|_\rho$ is equivalent to the usual sup-norm on $C([0, T]; \mathbb{R}^n)$. (Exercise: prove this.) And $C([0, T]; \mathbb{R}^n)$ is still a Banach space with the norm $|\cdot|_\rho$ for any $\rho > 0$. For short, *we will denote the Banach space $C([0, T]; \mathbb{R}^n)$ with norm $|\cdot|_\rho$ as C_ρ.*

Now define successive approximations $x_n(t)$ in C_ρ: put $x_0(t) = x_0$, the constant function, and for every $n \in \mathbb{N}$ and $t \in [0,T]$ set

$$x_n(t) = x_0 + \int_0^t f(s, x_{n-1}(s))\, ds\ . \tag{2.4.15}$$

If we define the (nonlinear) mapping $F : C_\rho \to C_\rho$ by

$$Fx(t) \ = \ x_0 + \int_0^t f(s, x(s))\, ds$$

equation (2.4.15) is $x_n(t) = Fx_{n-1}(t)$. It follows from the fundamental theorem of calculus that a function $x(\cdot) \in C^1([0,T]; \mathbb{R}^n)$ satisfies (2.4.13) and (2.4.14) if and only if it satisfies $x(t) = Fx(t)$, i.e., it is a fixed point of F.

Let's check that F maps C_ρ into C_ρ. The right side of (2.4.15) is clearly continuous (Exercise 2.4.20). Assume then that $|x|_\rho < \infty$. This means that, for some $C_1 > 0$,

$$|x(t)| \le C_1\, e^{\rho t}$$

for all $0 \le t \le T$. Set $y \equiv 0$ in (2.4.12) to obtain

$$|f(s,x(s))| \ \le \ |f(s,x(s)) - f(s,0))| + |f(s,0)| \ \le \ C_0|x(s)| + C_2 \ \le \ C_0C_1e^{\rho s} + C_2 \tag{2.4.16}$$

for a constant $C_2 = \sup_{0\le s\le T} |f(s,0)| < \infty$. We then have

$$\begin{aligned} |x_n(t)| &\le |x_0| + \int_0^t |f(s, x_{n-1}(s))|\, ds \\ &\le |x_0| + C_0C_1 \int_0^t e^{\rho s}\, ds + C_2\, t \\ &= |x_0| + C_0C_1(e^{\rho t} - 1)/\rho + C_2\, t \le C_3\, e^{\rho t} \end{aligned}$$

where $C_3 = |x_0| + C_0C_1/\rho + C_2$.

Finally we show that F is a contraction mapping on C_ρ provided ρ is sufficiently large. Take any $x, y \in C_\rho$. Then

$$\begin{aligned} |Fx(\cdot) - Fy(\cdot)|_\rho \ &= \ \sup_{0\le t\le T} e^{-\rho t} \Big| \int_0^t f(s,x(s)) - f(s,y(s))\, ds \Big| \\ &\le \ \sup_t \ e^{-\rho t} \int_0^t |f(s,x(s)) - f(s,y(s))|\, ds \\ &\le \ C_0 \sup_t \ e^{-\rho t} \int_0^t |x(s) - y(s)|\, ds \\ &= \ C_0 \sup_t \ e^{-\rho t} \int_0^t e^{\rho s}\ e^{-\rho s} |x(s) - y(s)|\, ds \\ &\le \ C_0 \sup_t \ e^{-\rho t} \int_0^t e^{\rho s} \sup_{0\le s\le t} \{e^{-\rho s}|x(s) - y(s)|\}\, ds \\ &= \ C_0 \sup_t \ e^{-\rho t} \int_0^t e^{\rho s}\, ds \ \sup_{0\le s\le t} \{e^{-\rho s}|x(s) - y(s)|\} \\ &= \ C_0 \sup_t \ \frac{1 - e^{-\rho t}}{\rho} \ \sup_{0\le s\le T} e^{-\rho s}|x(s) - y(s)| \\ &= \ \frac{C_0}{\rho}\ (1 - e^{-\rho T})\ |x(s) - y(s)|_\rho \\ &\le \ \frac{C_0}{\rho}\ |x(s) - y(s)|_\rho. \end{aligned}$$

When C_ρ is chosen so that $\rho > C_0$, F is a contraction on C_ρ. We then conclude that there is a unique fixed point $x \in C_\rho$ of F.

Finaly we observe that the fixed point x satisfies $x(0) = x_0$, and since $x \in C([0,T];\mathbb{R}^n)$, $f(\cdot, x(\cdot)) \in C([0,T];\mathbb{R}^n)$ as well. In Exercise 2.4.20 the reader is asked to show that $x(t) = Fx(t)$ can be continuously differentiated with respect to t, so $x(\cdot) \in C^1([0,T];X)$. □

2.4.6 Corollary. *Let the assumptions of Theorem 2.4.5 hold for every $T > 0$. (The bound (2.4.12) must hold on $0 \le t < \infty$ with the same constant C_1.) Then the initial value problem (2.4.13)-(2.4.14) has a unique solution $x(\cdot) \in C^1([0,\infty);\mathbb{R}^n)$.*

Proof. If the corollary were not true there would have to be some point $t > 0$ at which it failed. By taking $T > t$ in Theorem 2.4.5 we get a contradiction. □

The Lipschitz continuity assumption says f can only grow linearly in x. Indeed, if there is a $c > 0$ such that $|f(x) - f(y)| \le c\,|x - y|$ for all $x, y \in \mathbb{R}^n$ then we have

$$|f(x)| \le |f(0)| + |f(x) - f(0)| \le |f(0)| + c\,|x - 0| = |f(0)| + c\,|x|$$

for all x.

2.4.7 Example. The function $f(t,x) = x^2$ is not Lipschitz continuous on $-\infty < x < \infty$. However it is Lipschitz continuous on compact subsets of $\mathbb{R}$. For $|x^2 - y^2| \le |x + y|\,|x - y| \le 2M\,|x - y|$ on the set $-M \le x \le M$. So we do not expect the ODE $\dot{x} = x^2$ to be globally solvable, that is to have a solution for all t when given any initial conditions.

This equation is 'separable' and can be easily solved by integrating the equivalent equation $\frac{dx}{x^2} = dt$. The function $x(t) = \frac{1}{c-t}$ is the general solution of this equation (c is the integration constant). If we are given an initial condition, $x(0) = x_0$ say, we obtain $x(t) = \frac{1}{1/x_0 - t}$. But this solution blows up as $t \uparrow 1/x_0$. This initial value problem has no solution on $[0,\infty)$ when $x(0) > 0$.

From Theorem 2.4.1 we conclude that a solution exists on $[-b,b]$ for any $b < \min\{r/C_0, 1/C_1\}$. (Any $a > 0$ will satisfy the hypotheses of the theorem.) If we use the initial value $x(0) = x_0 \in \mathbb{R}$, the Lipschitz constant for f is $C_1 = 2\,\max\{|x_0 - r|, |x_0 + r|\}$ on the interval $[x_0 - r, x_0 + r]$. Taking for definiteness $x_0 > 0$ we have $C_1 = 2\,(x_0 + r)$. The maximum of f on $[x_0 - r, x_0 + r]$ is $(x_0 + r)^2$, again taking $x_0 > 0$ for definiteness. We conclude that a solution exists on $[-b,b]$ for any $b < \min\{r/(x_0 + r)^2, 1/[2(x_0 + r)]\} = 1/(x_0 + r)\,\min\{r/(x_0 + r), 1/2\}$. If $r > x_0$ then the last minimum is $1/2$, and if $r < x_0$ the minimum is $r/(x_0 + r) < 1/2$. So we might as well take $r = x_0$ and get the largest possible minimum of $1/2$. This then allows any $b < 1/(4x_0) = 1/(4r)$ to be used in our local existence theorem. The larger the initial value x_0, the shorter is the guaranteed interval of existence. This is consistant with the known solutions above.

2.4.8 Remark. The theorems in this section apply equally well when the values x are local coordinates for a manifold. Often the application is to a system independent of t so that $\dot{x} = F(x)$; the function F is interpreted as a vector field on the manifold, and a solution curve $x(t)$ as a curve on the manifold whose tangent vector at each point equals the vector F at that point. The first proof in particular makes evident the local nature of the theorem; we can solve the differential equation in one coordinate patch, and then in an overlapping patch we can 'restart' the solution and continue it into the next patch.

2.4.9 Exercise. Let $a \in \mathbb{R}$, $a \ne 0$. Compute the first four successive approximations for the equation $\dot{x} = ax$ with initial condition $x(0) = 1$.

Compute the first three successive approximations for the system

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -a^2 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \qquad \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

2.4.10 Exercise. Conjecture the solution of each ODE in the previous exercise. Use the successive approximation formula and an induction argument to prove your conjecture.

2.4.11 Exercise. Calculate the first three successive approximations for the equation $\dot{x} = t^2(1 - x)$ with $x(0) = 0$. (Boyce and DiPrima, 5ed, p 92, Ex 1.)

2.4.12 Exercise. Conjecture the general n-th successive approximation for the ODE in the previous exercise. Use induction to prove your conjecture.

2.4.13 Exercise. Calculate the first three successive approximations for each of the following equations:

(a) $\dot{x} = t - x$ with $x(0) = 1$.

(b) $\dot{x} = x + t^3$ with $x(0) = 1$.

(c) $\dot{x} = (t+1)^2 x$ with $x(0) = 1$.

2.4.14 Exercise. Calculate the first three successive approximations for

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -t^2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \qquad \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

2.4.15 Exercise. Show that Theorem 2.4.1 remains true if we replace $\mathbb{R}^n$ by $\mathbb{C}^n$ in its statement.

2.4.16 Exercise. Show that Theorem 2.4.2 remains true if we replace $\mathbb{R}^n$ by $\mathbb{C}^n$ in its statement.

2.4.17 Exercise. Show that Theorem 2.4.5 remains true if we replace $\mathbb{R}^n$ by $\mathbb{C}^n$ in its statement.

2.4.18 Exercise. Let $T \in (0, \infty]$ and $f \in C([0,T); \mathbb{R}^n)$, and let $A(t)$ be an $n \times n$ matrix whose entries are also in $C([0,T); \mathbb{R}^n)$. What additional conditions on $A(t)$ and $f(t)$ are needed to ensure that the linear inital value problem

$$\dot{x}(t) = A(t)x(t) + f(t) \qquad \text{and} \qquad x(0) = x_0 \tag{2.4.17}$$

has a unique solution x in $C^1([0,T); \mathbb{R}^n)$?

2.4.19 Exercise. Consider a system of first order difference equations

$$x_m = Ax_{m-1} + f_m$$

for $m = 1, 2, ...$, where $x_0 \in \mathbb{R}^n$ is given as is the sequence of vectors $f_m \in \mathbb{R}^n$. Here the $n \times n$ matrix A is independent of m. What can you say about the existence of a solution (a sequence) to this problem? If there is one, is the solution unique?

2.4.20 Exercise. Show that the right side of (2.4.15) is continuously differentiable when f and x_{n-1} are continuous.

2.5 Linear Equations with Variable Coefficients

In this section we study the linear system $\dot{x}(t) = A(t)x(t) + f(t)$ in more detail.

2.5.1 Theorem (linear systems)**.** *Let $T > 0$, $A \in C([0,T]; \mathbb{R}^{n\times n})$, $f \in C([0,T]; \mathbb{R}^n)$, and $x_0 \in \mathbb{R}^n$. Then the initial value problem*

$$\dot{x} = A(t)x + f(t) \qquad \text{on } \ 0 \le t \le T, \qquad \text{and} \qquad x(0) = x_0 \tag{2.5.1}$$

has a unique solution $x(\cdot) \in C^1([0,T]; \mathbb{R}^n)$.

Proof. The function $t \mapsto M(t) = \|A(t)\|$ is continuous on $[0,T]$. For the operator norm $\|\cdot\|$ is continuous on $\mathbb{R}^{n\times n}$ (see the dicussion following Theorem 1.1.10), so $M(\cdot)$ is a continuous function on the compact set $[0,T]$. The function $g(t,x) = A(t)x + f(t)$ on $[0,T] \times \mathbb{R}^n$ satisfies the hypothesis of the function f in Theorem 2.4.5. For the continuity of g is clear since A is continuous on $\mathbb{R}^n$ at each t, and the sum of the continuous vector functions $A(t)x$ and $f(t)$ is continuous on $[0,T] \times \mathbb{R}^n$. And it is easy to verify the Lipschitz continuity on $\mathbb{R}^n$:

$$|g(t,x) - g(t,y)| = |A(t)x - A(t)y| \le M(t)\,|x-y| \le C\,|x-y|\,.$$

We conclude from Theorem 2.4.5 that (2.5.1) has a unique solution in $C^1([0,T]; \mathbb{R}^n)$. □

2.5.2 Corollary. *For every $T > 0$ let the hypotheses of Theorem 2.5.1 hold. Here we allow the operator norm $M(t) = \|A(t)\|$ to become unbounded as $t \to \infty$, although it is necessarily bounded on every compact subinterval of $[0, \infty)$. Then there exists a unique solution $x(\cdot) \in C^1([0,\infty); \mathbb{R}^n)$ of the initial value problem*

$$\dot{x} = A(t)x + f(t) \qquad \text{on } \ t > 0, \qquad \text{and} \qquad x(0) = x_0 \ . \tag{2.5.2}$$

Proof. For every $T > 0$ Theorem 2.5.1 shows that there exists a unique solution to (2.5.2) in $C^1([0,T]; \mathbb{R}^n)$. If the corollary were not true there would be some $t > 0$ at which the existence of a unique C^1 solution fails. By taking $T > t$ in Theorem 2.5.1 we reach a contradiction. □

2.5.3 Definition. Let $T \in (0, \infty]$ and $A \in C([0,T); \mathbb{R}^{n\times n})$. A $\mathbb{R}^{n\times n}$-valued continuous function $W \in C([0,T), \mathbb{R}^{n\times n})$ which satisfies

$$\frac{d}{dt}W = A(t)W \qquad \text{on } \ 0 < t < T\,, \tag{2.5.3}$$

and is non-singular for every $t \in [0,T)$, is called a *fundumental solution* of the ODE $\dot{x} = A(t)x$ on $0 \le t < T$. If in addition $W(0) = I$, the $n \times n$ identity matrix, then we call W a *principle solution.*

The equation (2.5.3) means that each column of $W(t)$ satisfies the equation $\dot{x} = A(t)x$.

If $W(t)$ is a fundumental solution, then $W(t)W(0)^{-1}$ is a principle solution. If $W(t)$ is any fundumental solution we sometimes write

$$W(t,s) = W(t)W(s)^{-1}$$

for the 'principle solution' satisfying $W(s,s) = I$ when $t = s$. Obviously the function $W(t,s)$ satisfies the relation

$$W(t,s)\,W(s,r) \ = \ W(t,r) \tag{2.5.4}$$

whenever $0 \le r < s < t$.

2.5.4 Theorem (fundamental solutions)**.** *Let $T > 0$ and $A \in C([0,T]; \mathbb{R}^{n\times n})$. Then a fundamental solution $W(t)$ satisfying (2.5.3) exists on $0 \le t \le T$. Furthermore, this solution is unique if a non-singular initial matrix $W(0) = W_0$ is specified. In particular, a unique principle solution exists on $0 \le t \le T$.*

If $A \in C([0,\infty); \mathbb{R}^{n\times n})$ ($\|A(t)\|$ need not be bounded on $[0,\infty)$) then $W(t)$ exists for all $t \in [0,\infty)$. This W is unique if an initial condition $W(0) = W_0$ is specified.

Proof. Set $W(t) = [w_1(t)\, w_2(t) \ldots w_n(t)]$ where $w_j(t)$ is the j-th column of $W(t)$. Equation (2.5.3) means that $\dot{w}_j = A(t)w_j$ for every j. And the initial condition $W(0) = W_0$ means that $w_j(0)$ equations the j-th column of W_0. Theorem 2.4.5 then implies that, for each j, there is a unique function $w_j(t)$ satisfying these conditions. Whence, $W(t)$ satisfies (2.5.3) and $W(0) = W_0$.

Next we show that $W(t)$ is non-singular for all t if W_0 is non-singular. From Corollary 1.3.20 we know that if the inverse $W^{-1}(t)$ exists, this inverse also exists on some small open interval about t. We may then differentiate the expression $WW^{-1} = I$ with respect to t to obtain

$$\frac{d}{dt}W^{-1}(t) = -W^{-1}(t)\frac{d}{dt}[W(t)]W^{-1}(t).$$

Inserting the right side of (2.5.3) for $\frac{d}{dt}W$ in this equation gives

$$\frac{d}{dt}W^{-1}(t) = -W^{-1}(t)A(t).$$

Now the equation

$$\frac{d}{dt}V = -VA(t) \qquad \text{on } \ 0 < t < T \tag{2.5.5}$$

can be analyzed in the same way as (2.5.3); it has a unique solution on $[0,T]$ for any $V_0 \in \mathbb{R}^{n\times n}$. (One may in fact take the transpose of (2.5.5) to see it can be put into the same form as (2.5.3).) Therefore if we set $V(0) = W_0^{-1}$, the unique solution of (2.5.5) gives us $W^{-1}(t)$ for any t in $[0,T)$. □

2.5.5 Corollary. *For every $T > 0$ let the hypotheses of Theorem 2.5.4 hold. Here we allow the operator norm $M(t) = \|A(t)\|$ to become unbounded as $t \to \infty$, although it is necessarily bounded on every compact subinterval of $[0, \infty)$. Then there exists a unique solution $W(\cdot) \in C^1([0,\infty), \mathbb{R}^{n\times n})$ of the initial value problem*

$$\dot{W} \;=\; A(t)W \qquad \text{on} \;\; t > 0, \qquad\qquad \text{and} \qquad\qquad W(0) \;=\; W_0 \;. \tag{2.5.6}$$

Further, $W(t)^{-1}$ exists for every $t \geq 0$ if W_0 is non-singular.

The proof is the same as the proof of Corollary 2.5.2.

2.5.6 Theorem (variation of parameters formula)**.** *If W is a principle solution of $\dot{x} = A(t)x$ on $0 \leq t < T$ then the (unique) solution of the initial value problem*

$$\dot{x} = A(t)x + f(t) \qquad \text{on} \;\; 0 < t < T, \qquad\qquad \text{and} \qquad\qquad x(0) = x_0 \tag{2.5.7}$$

is given by

$$x(t) \;=\; W(t)\left[x_0 + \int_0^t W(s)^{-1} f(s)\, ds\right] \;=\; W(t)x_0 + \int_0^t W(t,s) f(s)\, ds \;, \tag{2.5.8}$$

where $W(t,s) = W(t)W^{-1}(s)$.

Proof. Multiply (2.5.7) on the left by $W^{-1}(t)$ to obtain

$$W^{-1}(t)\dot{x}(t) - W^{-1}(t)A(t)x(t) = W^{-1}(t)f(t).$$

In this equation W^{-1} functions as an integrating factor as the following computation shows:

$$\begin{aligned}\frac{d}{dt}[W^{-1}(t)x(t)] &= W^{-1}(t)\dot{x}(t) - W^{-1}(t)\dot{W}(t)W^{-1}(t)x(t)\\ &= W^{-1}(t)\dot{x}(t) - W^{-1}(t)A(t)x(t).\end{aligned}$$

Thus,

$$\frac{d}{dt}[W^{-1}(t)x(t)] = W^{-1}(t)f(t)$$

which may be written

$$W^{-1}(t)x(t) = x(0) + \int_0^t W^{-1}(s)f(s)\, ds$$

since $W(0) = I$. Now the formula (2.5.8) is evident.[16] □

2.5.7 Exercise. Consider the linear constant coefficient equation $y'' + ay' + by = f(t)$. Give the homogeneous ('complimentary') solution, the Wronskian, and the particular solution using the variation of parameters formula given in an introductory ODE text. (Assume $y(0)$ and $y'(0)$ are given.)

Turn this equation into a 2×2 system and compute $W(t)$, and the solution of the system using (2.5.8). Compare answers.

2.5.8 Exercise. Let $A \in C([0,T]; \mathbb{R}^{n\times n})$ and $x(t)$ be a solution of $\dot{x} = A(t)x$ on $[0,T]$. Show that either $x(t) \neq 0$ for any $t \in [0,T)$, or else $x(t) = 0$ for all $t \in [0,T)$.

2.5.9 Exercise. Let $[a,b]$ be a compact interval in $\mathbb{R}$, $f \in C([a,b]; \mathbb{R}^n)$, $t_0 \in [a,b]$, and define $x(t) = \int_{t_0}^t f(s)\, ds$ for $t \in [a,b]$. Give a careful proof that x is continuous on $[a,b]$, and in particular, that the limits $x(b-)$ and $x(a+)$ exist. Prove these same facts about $x(t)$ if we only assume $f \in L^1((a,b); \mathbb{R}^n)$.

2.5.10 Exercise. Assume $f(t) \in L^1(0,T; \mathbb{R}^n)$ and $A(t) \in C([0,T]; \mathbb{R}^{n\times n})$. Show $x(t)$ given by (2.5.8) is continuous, and that $\dot{x}(t) \in L^1(0,T; \mathbb{R}^n)$.

2.5.11 Exercise. Let the assumptions of Theorem 2.5.4 hold. Give an alternate proof that $W(t)$ is non-singular for all $t \in [0,T]$ by completing the following outline. If $W(t)$ were singular at some $t_0 \in [0,T]$ then there would be an element $x_0 \in X$ such that $x_0 \neq 0$ but $W(t_0)x_0 = 0$. Now let $x(t)$ be the unique solution on $[0,T]$ of the initial value problem $\dot{x} = A(t)x$ with $x(0) = W(0)x_0 = 0$. Use the fact that $x(t) \equiv 0$ is a solution to this initial value problem to conclude that $W(t)$ is singular for all $t \in [0,T]$.

[16] TO STEVE: Include the theorem about the Wronskian and trace.

2.6 Newtonian Mechanics and Electric Circuits

Newton's equations of motion, and equations for systems of electric circuits, are often written in terms of the second time derivative of a position, or charge, vector $x(t)$. We will give a few results for equations explicitly in this form.

2.6.1 Theorem (general existence)**.** *Let $T > 0$ and $f : [0,T] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ be continuous. And assume there is a constant $C_0 > 0$ such that*

$$|f(t,x_1,y_1) - f(t,x_2,y_2)| \;\le\; C_0\,(\,|x_1 - x_2| + |y_1 - y_2|\,)$$

for all $t \in [0,T]$ and $x_1, x_2, y_1, y_2 \in \mathbb{R}^n$. And assume the 'mass' matrix M of dimension $n \times n$ is non-singular (and constant in t). Then the initial value problem

$$M\ddot{x} \;=\; f(t,x,\dot{x}) \tag{2.6.1}$$

with $x(0) = x_0$ and $\dot{x}(0) = x_1$, has a unique solution $x(\cdot) \in C^2([0,T];\mathbb{R}^n)$ for any $x_0, x_1 \in \mathbb{R}^n$.

Proof. Set $y(t) = \dot{x}(t)$ so that $\dot{y} = M^{-1}\,f(t,x,y)$. Then (2.6.1) is equivalent to the first order system

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} \;=\; \begin{pmatrix} y \\ M^{-1}\,f(t,x,y) \end{pmatrix}\;. \tag{2.6.2}$$

The initial conditions for this system are $x(0) = x_0$ and $y(0) = x_1$. If we set $F(t,x,y)$ equal to the right side of (2.6.2), F is a $2n$-vector valued function which is continuous on $[0,T] \times \mathbb{R}^{2n}$, and which satisfies:

$$\begin{aligned}
|F(t,x_1,y_1) - F(t,x_2,y_2)| \;&=\; \left|\begin{pmatrix} y_1 - y_2 \\ M^{-1}[f(t,x_1,y_1) - f(t,x_2,y_2)] \end{pmatrix}\right| \\
&\le\; |y_1 - y_2| + \|M^{-1}\|\;|f(t,x_1,y_1) - f(t,x_2,y_2)| \\
&\le\; |y_1 - y_2| + C_0\,\|M^{-1}\|\;(\,|x_1 - x_2| + |y_1 - y_2|\,) \\
&\le\; (1 + C_0\,\|M^{-1}\|\,)\;(\,|x_1 - x_2| + |y_1 - y_2|\,) \\
&\le\; (1 + C_0\,\|M^{-1}\|\,)\;\sqrt{|x_1 - x_2|^2 + |y_1 - y_2|^2}\;.
\end{aligned}$$

Here we have used the inequalities $\sqrt{|x|^2 + |y|^2} \le |x| + |y|$ and $|x| + |y| \le \sqrt{2}\sqrt{|x|^2 + |y|^2}$. The second is a result of $2ab \le a^2 + b^2$, whence $(a+b)^2 \le 2(a^2 + b^2)$, for all $a, b \ge 0$.

This shows that F satisfies the Lipschitz continuity hypothesis of Theorem 2.4.5, and this theorem follows from that one. We have $x(\cdot) \in C^2$ because Theorem 2.4.5 shows that $\dot{x}$ as well as x is in C^1. □

2.6.2 Corollary. *Let M, R, and K be $n \times n$ (constant) matrices with M non-singular, and let $f : [0,T] \to \mathbb{R}^n$ be continuous. Then the linear, constant coefficient mechanical (electrical) system*

$$M\ddot{x} + R\dot{x} + Kx \;=\; f(t)\;, \tag{2.6.3}$$

with mass (inductance) M, resistence R, and stiffness (inverse capacitance) K, has a unique solution $x(\cdot) \in C^2([0,T];\mathbb{R}^n)$ when given any initial conditions $x(0)$ and $\dot{x}(0)$ in $\mathbb{R}^n$.

In fact the corollary is true even if f has jump discontinuities at any discrete set of times in the interval $(0,T)$, except that x then may only be C^1 at those points of discontinuity.

2.6.3 Example. Let $x \in \mathbb{R}^{3n}$ denote the positions of n particles in $\mathbb{R}^3$ which are moving in a force field given by $-\nabla q(x)$ where $q(x) = \frac{1}{2}x^t Q x$ is a potential given by the quadratic form, and Q is symmetric and positive definite. Then x satisfies

$$M\ddot{x} \;=\; -\nabla q(x) \;=\; Qx$$

where M is a diagonal matrix of (strictly positive) masses for each particle.

2.6.4 Example. Suppose we wish to solve the wave equation

$$u_{tt}(t,x) \;=\; \Delta u(t,x)$$

when $x \in \Omega$, a bounded region in $\mathbb{R}^d$. We also require that $u = 0$ on $\partial\Omega$. To numerically approximate u, we take n 'basis' or 'shape' functions $w_j(x)$ in Ω, $j = 1, 2, ..., n$, (often n must be fairly large to get a good approximation) and write

$$u(t,x) \;\approx\; \sum_{j=1}^{n} c_j(t)\, w_j(x)$$

where the coefficients c_j will change with time. We require that each $w_j = 0$ on $\partial\Omega$. The 'approximate wave equation' would then be

$$\sum_{j=1}^{n} \ddot{c}_j(t)\, w_j(x) \;=\; \sum_{j=1}^{n} c_j(t)\, \Delta w_j(x)\ .$$

It is too much to ask that this equation hold at evey $x \in \Omega$ and still be able to solve for the n unknown $c_j(t)$'s. Instead we only ask that this equation hold when 'projected' onto an n dimensional subspace of functions spanned by the w_j's. That is, we only require that

$$\int_\Omega u_{tt}(t,x)\, w_i(x)\, dx \;=\; \int_\Omega \Delta u(t,x)\, w_i(x)\, dx$$

hold for $i = 1, ..., n$. With the approximation inserted this becomes

$$\sum_{j=1}^{n} \ddot{c}_j(t) \int_\Omega w_j(x)\, w_i(x)\, dx \;=\; \sum_{j=1}^{n} c_j(t) \int_\Omega \Delta w_j(x)\, w_i(x)\, dx \;=\; -\sum_{j=1}^{n} c_j(t) \int_\Omega \nabla w_j(x)\cdot\nabla w_i(x)\, dx$$

where the last equality results from integrating each term

$$\frac{\partial^2 w_j}{\partial x_k^2}\, w_i$$

by parts one time and using the zero boundary conditions.

If we let M denote the matrix with ij-th entry $\int_\Omega w_j(x)\, w_i(x)\, dx$, and let K denote the matrix with ij-th entry $\int_\Omega \nabla w_j(x)\cdot\nabla w_i(x)\, dx$, this system of equations can be written

$$M\ddot{\mathbf{c}} \;=\; -K\mathbf{c}$$

where $\mathbf{c} = \mathbf{c}(t)$ is the vector function with components $c_j(t)$. When inital conditions $\mathbf{c}(0)$ and $\dot{\mathbf{c}}(0)$ are given (corresponding of course to $u(0,x)$ and $u_t(0,x)$) the preceding corollary shows that this system has a unique solution. The above procedure is known as the *method of finite elements* or *Galarkin's approximation.*

2.6.5 Theorem. *Assume the real $n \times n$ matrices M and K are symmetric and positive definite, and that $f(t)$ is integrable on any compact subset of $[0, \infty)$. Then the linear constant coefficient equation*

$$M\ddot{x} + Kx \;=\; f(t) \tag{2.6.4}$$

has a general solution of the form $x(t) = x_c(t) + x_p(t)$ where a particular solution is

$$x_p(t) \;=\; M^{-1/2} \sin(At)\, A^{-1} \int_0^t \cos(As)\, M^{-1/2}\, f(s)\, ds - M^{-1/2} \cos(At)\, A^{-1} \int_0^t \sin(As)\, M^{-1/2}\, f(s)\, ds \tag{2.6.5}$$

and the complimentary solution is

$$x_c(t) \;=\; M^{-1/2} \cos(At)\, M^{1/2}\, x_0 + M^{-1/2} \sin(At)\, A^{-1}\, M^{1/2}\, x_1\ . \tag{2.6.6}$$

*Here x_0 and x_1 are arbitrary vectors in $\mathbb{R}^n$. The solution $x(t)$ given by the sum of (2.6.5) and (**??**) satisfies $x(0) = x_0$ and $\dot{x}(0) = x_1$.*

Proof. We will only sketch the proof, leaving the details to the reader. Since M is symmetric and positive definite it has a square root and (2.6.4) is the same as

$$\ddot{y} + A^2\, y \;=\; \tilde{f}(t) \tag{2.6.7}$$

where $y(t) = M^{1/2}\, x(t)$, $A^2 = \tilde{K} = M^{-1/2}\, K\, M^{-1/2}$ for some symmetric positive definite A, and $\tilde{f}(t) = M^{-1/2}\, f(t)$. So we solve (2.6.7); the reader can transform back to the $x(t)$ coordinates at the end.

The complimentary solution

$$y_c(t) \;=\; \sin(At)\, y_1 + \cos(At)\, y_0 \;,$$

which solves $\ddot{y} + A^2\, y = 0$, was obtained in Exercise 2.1.21.

Finding the particular solution can be done using the same trick as used for second order scalar equations. (It can also be done by transforming equation (2.6.7) into a first order $2n \times 2n$ system and applying the formulas (variation of parameters) developed in section 2.1.) We assume that y_p has the form

$$y_p(t) \;=\; \sin(At)\, u_1(t) + \cos(At)\, u_2(t) \tag{2.6.8}$$

for some unknown functions u_1 and u_2. (This is the same form as the complimentary solution except that the constants y_0 and y_1 of integration have been replaced by unknown functions.) We now plug (2.6.8) into (2.6.7) to solve for u_1 and u_2. The computation is simplified by the (perhaps unmotivated) assumption that

$$\sin(At)\, \dot{u}_1(t) + \cos(At)\, \dot{u}_2(t) \;=\; 0 \;. \tag{2.6.9}$$

We obtain

$$\cos(At)\, \dot{u}_1(t) - \sin(At)\, \dot{u}_2(t) \;=\; A^{-1}\, \tilde{f}(t) \;. \tag{2.6.10}$$

If we combine (2.6.9) and (2.6.10), and use Exercise 2.1.18, we arrive at

$$\begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix} \;=\; \int_0^t \begin{pmatrix} A^{-1}\, \cos(As)\, \tilde{f}(s) \\ -A^{-1}\, \sin(As)\, \tilde{f}(s) \end{pmatrix} ds \;.$$

One can directly verify that the resulting (2.6.8) does indeed solve (2.6.7). □

3 Partial Differential Equations

This chapter is devoted to solving the baisc boundary and initial value problems that arise in physics and engineering. Emphasis is on the self-adjoint case; then an eigen-decomposition holds and gives insight into the structure of solutions and formulas for the parabolic and hyperbolic equations.

3.1 Sturm-Liouville Theory

See section 4.2 in Dettman, *Mathematical Methods in Physics and Engineering.*

Also, section 7.5 in Naylor and Sell, *Linear Operators in Science and Engineering*, is excellent.

3.1.1 Exercise. TBD.

3.2 Separation of Variables

See sections 4.1, 4.3, and 4.4 in Dettman, *Mathematical Methods in Physics and Engineering.*

Many introductory books on partial differential equation present this method, including chapter 12 in Kreyszig, *Advanced Engineering Mathematics* and sections 7.8 and 7.9 in Naylor and Sell, *Linear Operators in Science and Engineering.* An especially thorough presentiation is given by R. Haberman, *Applied Partial Differential Equations.*

Physics books on electricity and magnetism also give many examples of this method.

3.2.1 Exercise. TBD.

3.3 Finite Element or Galerkin Approximations

3.3.1 Exercise. TBD.

3.4 Weak Theory of Elliptic Boundary Value Problems

Recall notation $\langle , \rangle$, $\langle , \rangle_0$, $\langle , \rangle_1$, $\langle , \rangle_{-1}$, etc, and spaces of functions, including $L^2(0,T;L^2(\Omega))$, etc.

Now let's recall a few mathematical facts.

The *Riesz representation theorem*: If H is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$, then for every bounded linear functional $f : H \to \mathbb{C}$ there is a unique $u \in H$ which satisfies

$$\langle v, u \rangle \;=\; f(v) \qquad \text{for all } v \in H\,.$$

Moreover, the operator norm of f, $\sup\{|f(v)| \;;\; \|v\| = 1\}$, equals the Hilbert space norm $\|u\|$ of u.

The *Lax-Milgram lemma*: If H is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$, and if $B : H \times H \to \mathbb{C}$ is a sesqui-linear form which is

bounded: there is a $C > 0$ such that $\quad |B(u,v)| \le C\,\|u\|\,\|v\| \quad$ for all $u, v \in H$, and

coercive: there is a $c > 0$ such that $\quad |B(u,u)| \ge c\,\|u\|^2 \quad$ for all $u \in H$,

then for every bounded linear functional $f : H \to \mathbb{C}$ there is a unique $u \in H$ which satisfies

$$B(v, u) \;=\; f(v) \qquad \text{for all } v \in H\,.$$

Fredholm's theorem: If H is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, T is a compact operator on H with Hilbert space adjoint T^*, and for each $\lambda \in \mathbb{C}$ we define the null spaces

$$\begin{aligned} N_\lambda \;&=\; \mathcal{N}(T - \lambda) \;=\; \{w \in H \;;\; Tw = \lambda w\}\,, \quad \text{and} \\ N_\lambda^* \;&=\; \mathcal{N}(T^* - \lambda) \;=\; \{w \in H \;;\; T^* w = \lambda w\}\,, \end{aligned}$$

then

a) the set of $\lambda \in \mathbb{C}$ for which $N_\lambda \neq \{0\}$ is finite or countable, and if countable the only limit point of these λ's is 0;

b) if $\lambda \neq 0$ in $\mathbb{C}$, $\dim(N_\lambda) = \dim(N_{\overline{\lambda}}^*) < \infty$;
c) if $\lambda \neq 0$ in $\mathbb{C}$, the range space $\mathscr{R}(T-\lambda)$ is closed in H.

The *Rellich embedding theorem*: For any bounded $\Omega \subset \mathbb{R}^d$, the enbedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$ is a compact linear transformation. The same is true of $H_0^1(\Omega)$, or any other subspace of $H^1(\Omega)$.

The *divergence theorem and integration by parts*: If $\vec{f}$ is a vector field defined in a neighborhood of a bounded open set $\Omega \subset \mathbb{R}^d$ and $\vec{n}$ is the outward pointing normal to the boundary $\partial\Omega$, we have

$$\int_\Omega \nabla \cdot \vec{f}\, dx \;=\; \int_{\partial\Omega} \vec{n} \cdot \vec{f}\, ds\,.$$

(This is a multivariate version of the fundamental theorem of calculus.) If we combine this formula with the *product rule*

$$\nabla \cdot (v\, \nabla u) \;=\; \nabla v \cdot \nabla u + v\, \Delta u\,,$$

where u and v are scalar functions on Ω, we obtain an integration by parts formula

$$\int_\Omega v\, \Delta u\, dx \;=\; -\int_\Omega \nabla v \cdot \nabla u\, dx + \int_{\partial\Omega} v\, \frac{\partial u}{\partial n}\, ds$$

where we use the notation $\frac{\partial u}{\partial n} = \vec{n} \cdot \nabla u$. (Here we have set $\vec{f} = v\, \nabla u$.)

Weak derivatives: The weak theory of elliptic boundary problems depends critically on the notion of weak (or distributional) derivatives. We must take derivatives of functions which are not differentiable in the usual sense (limits of difference quotients existing). So an extension of the notion of derivative is needed, an extension which must be consistant with the elementary definition.

If $\Omega \subset \mathbb{R}^d$ is open and $f \in L^2(\Omega)$, we define a linear functional $\partial_i f : H_0^1(\Omega) \to \mathbb{C}$, called the *weak partial derivative* of f, by the formula

$$\partial_i f(v) = -\int_\Omega f\, \partial_i v\, dx\,.$$

Notice that if f is in $C^1(\Omega)$ and $v \in C_0^1(\Omega)$, this formula would hold because we could integrate the x_i variable by parts. We may abuse notation and write the left-hand side of this equation as $\int_\Omega \partial_i f\, v\, dx$ even if f is only in L^2.

The weak partial derivative of $f \in L^2(\Omega)$ is bounded on $H_0^1(\Omega)$ because

$$|-\int_\Omega f\, \partial_i v\, dx| \;\leq\; \Big(\int_\Omega |f|^2\, dx\Big)^{1/2} \Big(\int_\Omega |\partial_i v|^2\, dx\Big)^{1/2} \;\leq\; \|f\|_0\, \|v\|_1$$

thanks to Schwarz's inequality.

Notice that the formula $\int_\Omega \partial_i f\, v\, dx \;=\; -\int_\Omega f\, \partial_i v\, dx$ would not necessarily hold if v was only in $C^1(\overline{\Omega})$ because the integration by parts, using the divergence theorem with $\vec{f} = f v \vec{e}_i$ (the i-th unit basis vector), gives us

$$\int_\Omega \partial_i f\, v\, dx \;=\; -\int_\Omega f \partial_i v\, dx + \int_{\partial\Omega} n_i f v\, ds$$

where n_i is the i-th component of $\vec{n}$. Unless $v = 0$ on $\partial\Omega$, the last integral will not in general be zero.

3.4.1 Second order differential operators and Dirichlet forms

We will study linear, second order partial differential operators of the form[17]

$$Au \;=\; -\sum_{i,j=1}^d \partial_i(a_{ij}(x)\partial_j u) + \sum_{i=1}^d b_i(x)\partial_i u + c(x)u\,. \tag{3.4.1}$$

[17] Every second order linear operator,

$$A = \sum_{i,j=1}^d \tilde{a}_{ij}\partial_{ij} + \sum_1^d \tilde{b}_i\partial_i + \tilde{c}(x)$$

can be put into the form (3.4.1) if the coefficient functions are smooth enough. The reader can write out the translation from one form to the other. The form (3.4.1) is called the divergence form; it is better suited for integrating the highest order term by parts - using the divergence theorem.

We will assume the coefficient functions $a_{ij}(x), b_i(x), c(x)$ belong to $L^\infty(\Omega)$, the vector space of bounded (but not necessarily continuous), complex-valued functions on Ω.

In equation (3.4.1) the expression $\partial_i(a_{ij}(x)\partial_j u$ makes no sense in terms of the product rule from basic calculus unless each $a_{ij}(x) \in C^1(\Omega)$. But we will see that this expression can be given sense by moving the ∂_i onto a 'test function' v which is smooth enough to be differentiated using basic calculus. We will do this in such a way that it is consistent with the rules of calculus when $a_{ij}(x) \in C^1(\Omega)$. For any $v \in C_0^1(\Omega)$ or smoother, the correct formula is

$$-\int_\Omega \bar{v}\,\partial_i(a_{ij}(x)\partial_j u)\,dx \;=\; \int_\Omega \partial_i\bar{v}\,a_{ij}(x)\,\partial_j u\,dx. \tag{3.4.2}$$

When $a_{ij}(x) \in C^1(\Omega)$ this can be verified by integrating by parts, which we will do later in this section. When $a_{ij}(x) \in L^\infty(\Omega)$ this can be used as a *definition* of the left hand side.

3.4.1 Lemma. *Let $\Omega \subset \mathbb{R}^d$ be open and A be given in (3.4.1). Assume that the coefficient functions $a_{ij}(x)$, $b_i(x)$, and $c(x)$ belong to $L^\infty(\Omega)$. Then*

$$A : H^1(\Omega) \to H^{-1}(\Omega)$$

is a continuous linear transformation. In particular, the image of $H^1(\Omega)$ under A lies in $H^{-1}(\Omega)$, the dual vector space of $H_0^1(\Omega)$. This statement also holds for any subspace of $H^1(\Omega)$, for instance it holds for $H_0^1(\Omega)$.

Proof. The proof is performed in three steps:

a) Show $\partial_j : H^1(\Omega) \to L^2(\Omega)$ is continuous.

b) Show $\partial_j : L^2(\Omega) \to H^{-1}(\Omega)$ is continuous. (This is a 'weak' derivative.)

c) Show that multiplication by an $L^\infty(\Omega)$ function $a(x)$ is continuous, $a : L^2(\Omega) \to L^2(\Omega)$.

The terms of the differential operator A are made of sums and compositions of the above continuous operations. For instance the operator $\partial_i(a(x)\partial_j u)$ first maps the H^1 function u into its partial derivative $\partial_j u$ in L^2, then multiplies by $a(x)$ taking an L^2 function to another L^2 function, and finally 'differentiates' this L^2 function into an H^{-1} 'function'. Every operation is continuous.

a) The operator $\partial_j : H^1 \to L^2$ is continuous (bounded) since

$$\|\partial_j u\|_0^2 \le \|u\|_0^2 + \sum_{i=1}^d \|\partial_i u\|_0^2 = \|u\|_1^2\,.$$

We have a simple bound on the operator norm: $\|\partial_j\| \le 1$.

b) For $u \in L^2$, the weak derivative $\partial_j u$ is defined by

$$\langle \partial_j u, v\rangle \;=\; -\int_\Omega u\,\partial_j v\,dx$$

when $v \in H_0^1(\Omega)$. Using Schwarz's inequality,

$$|-\int_\Omega u'\partial_j v\,dx| \;\le\; \big(\int_\Omega |u|^2\,dx\big)^{1/2}\,\big(\int_\Omega |\partial_j v|^2\,dx\big)^{1/2} \;\le\; \|u\|_0\,\|v\|_1\,,$$

where the last inequality was also used in part (a). Thus $\partial_j : L^2 \to H^{-1}$.

This transformation is continuous, for

$$\begin{aligned}
\|\partial_j u\|_{-1} \;&=\; \sup_{v\in H_0^1,\,\|v\|_1=1} |\langle \partial_j u, v\rangle| \;=\; \sup_{v\in C_0^\infty,\,\|v\|_1=1} |\langle \partial_j u, v\rangle| \\
&=\; \sup_{v\in C_0^\infty,\,\|v\|_1=1} |-\int_\Omega u'\partial_j v\,dx| \;\le\; \sup_{v\in C_0^\infty,\,\|v\|_1=1} \Big(\int_\Omega |u|^2\,dx \int_\Omega |\partial_j v|^2\,dx\Big)^{1/2} \\
&\le\; \sup_{v\in C_0^\infty,\,\|v\|_1=1} \big(\int_\Omega |u|^2\,dx\big)^{1/2}\,\|v\|_1 \;=\; \|u\|_0.
\end{aligned}$$

The first equality above is the definition of the norm in H^{-1}, and we recall that C_0^∞ is a dense subset of H_0^1. So the operator norm of ∂_j satisfies $\|\partial_j\| \le 1$.

c) Finally, let $a(x) \in L^\infty$ and define $Mu(x) = a(x)\,u(x)$ when $u \in L^2$. We have

$$\|a\,u\|_0 \;=\; \big(\int_\Omega |au|^2\,dx\big)^{1/2} \;\le\; \sup_{x\in\Omega}|a(x)|\,\big(\int_\Omega |u|^2\,dx\big)^{1/2} \;=\; \sup_{x\in\Omega}|a(x)|\,\|u\|_0$$

which shows that the operator norm $\|M\| \le \sup_{x\in\Omega}|a(x)|$. □

A sesqui-linear form,[18] or *Dirichlet form*, corresponding to (3.4.1) is

$$B(u,v) = \sum_{i=1}^d\sum_{j=1}^d \int_\Omega \partial_i\bar v\, a_{ij}(x)\partial_j u\,dx + \sum_{i=1}^d \int_\Omega \bar v\, b_i(x)\partial_i u\,dx + \int_\Omega \bar v\, c(x)u\,dx. \tag{3.4.3}$$

The basic relation between the differential equation and its Dirichlet form is given in the next result.

3.4.2 Lemma. *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set and A be given in (3.4.1). Assume that the coefficient functions $a_{ij}(x)$, $b_i(x)$, and $c(x)$ belong to $L^\infty(\Omega)$. Then the equation*

$$\langle Au, \bar v\rangle \;=\; \Big\langle -\sum_{i,j=1}^d \partial_i(a_{ij}(x)\partial_j u) + \sum_{i=1}^d b_i(x)\partial_i u + c(x)u\,,\,\bar v\Big\rangle \;=\; B(u,v) \tag{3.4.4}$$

holds for all $u, v \in H_0^1(\Omega)$.

Proof. By Lemma 3.4.1 the sesqui-linear mapping

$$H_0^1 \times H_0^1 \ni (u,v) \mapsto \langle Au, \bar v\rangle \in \mathbb{C}$$

is continuous in both arguments. It therefore suffices to show (3.4.4) for $u, v \in C_0^\infty(\Omega)$.[19] For these functions the duality pairing is $\langle u, v\rangle = \int_\Omega vu\,dx$. We integrate by parts the highest order derivative term and use the fact that both u and v are zero near $\partial\Omega$:

$$-\int_\Omega \bar v \sum_{i,j=1}^d \partial_i(a_{ij}(x)\partial_j u)\,dx \;=\; -\sum_{i,j}\int_\Omega \bar v\,\partial_i(a_{ij}(x)\partial_j u)\,dx \;=\; \sum_{i,j}\int_\Omega (\partial_i\bar v)\,a_{ij}(x)\partial_j u\,dx\,.$$

The equality of the other terms on the left side of (3.4.4) and the right side of (3.4.3) is clear. □

3.4.3 Exercise. Let $a, b \in \mathbb{R}$. Show that $ab \le \frac12(a^2+b^2)$. (Notice that $0 \le (a-b)^2$.) Let $\epsilon > 0$; show also that $ab \le \frac12(\epsilon\,a^2 + b^2/\epsilon)$. Let f and g be functions in $L^2(\Omega)$ and let $\|f\|_0^2 = \int_\Omega |f|^2\,dx$. Show that for every $\epsilon > 0$,

$$|\int_\Omega fg\,dx| \;\le\; \frac12\big(\epsilon\|f\|_0^2 + \frac{1}{\epsilon}\|g\|_0^2\big)\,.$$

3.4.4 Exercise. Let $u \in H_0^1(\Omega)$. Show that $\|u\|_0 \le \|u\|_1$, and that $\|\partial_i u\|_0 \le \|u\|_1$. Show that $H_0^1(\Omega) \subset L^2(\Omega)$, and that $v \mapsto \int_\Omega fv\,dx$ is a bounded linear functional on $H_0^1(\Omega)$ when $f \in L^2(\Omega)$.

[18] $B(u,v)$ is sesqui-linear if $B(\alpha u + \beta v, w) = \alpha B(u,w) + \beta B(v,w)$ and $B(u, \alpha v + \beta w) = \bar\alpha B(u,v) + \bar\beta B(u,w)$ for all functions u, v, w and scalars α, β.

[19] Clarification: For general $u, v \in H_0^1$ there is a sequence of C_0^∞ functions, u_n, v_n, which converge in H_0^1 to u and v. (This is because C_0^∞ is dense in H_0^1.) Now the continuity of $\langle Au, v\rangle$ in both u and v means precisely that we can interchange the following limits

$$\langle Au, v\rangle = \langle A\lim_m u_m, \lim_n v_n\rangle = \lim_m\lim_n \langle Au_m, v_n\rangle.$$

Thus, if (3.4.4) holds for all C_0^∞ functions u_m and v_n, then passing to the limit on both sides of (3.4.4) shows that this same equation holds also for all $u, v \in H_0^1$.

3.4.5 Exercise. Show that the differential operator (3.4.1) can be written in the form

$$\sum_{i,j=1}^{d} \tilde{a}_{ij}\partial_{ij} + \sum_{1}^{d} \tilde{b}_i\partial_i + \tilde{c}(x)\,.$$

Give the formulas to convert the coefficient functions of one into the other.

3.4.6 Exercise. Show $B(u,v)$ is sesqui-linear.

3.4.7 Exercise. Let A be given by (3.4.1). Is it always true that $(u,v) \mapsto \int_\Omega \bar{v}\, Au\, dx$ is an inner product on $C_0^\infty(\Omega)$? What fails? Is it always true that $(u,v) \mapsto B(u,v)$ is an inner product on $C_0^\infty(\Omega)$?

3.4.8 Exercise. Extend the definition of weak derivative to $\partial^\alpha f$ for any $\alpha \in \mathbb{N}_0^d$, when $f \in L^2(\Omega)$ by the formula

$$\langle \partial^\alpha f, v\rangle \;=\; (-1)^{|\alpha|} \int_\Omega f\, \partial^\alpha v\, dx$$

when $v \in H_0^{|\alpha|}(\Omega)$. Show that this definition is consistent with the elementary definition of ∂^α when f and v are $C^{|\alpha|}(\Omega)$ functions and v has compact support. Show also that the operator $\partial^\alpha f$ is a *bounded* linear functional on $H_0^{|\alpha|}(\Omega)$.

3.4.9 Exercise. Assume $\Omega \subset \mathbb{R}^d$ is a bounded open set. Show that $L^2(\Omega) \subset L^1(\Omega)$ by applying Schwarz's inequality to $\int_\Omega fg\, dx$ with $g \equiv 1$. Is this result true when Ω is not bounded?

3.4.2 Elliptic operators and energy bounds

The following result is critical.

3.4.10 Lemma (energy inequality for elliptic equations)**.** *Assume $c_0 > 0$, $a_{ij}(x) \in L^\infty(\Omega)$, and*

$$c_0|z|^2 \;\le\; \Re\Big\{ \sum_{i,j=1}^{d} a_{ij}(x)\bar{z}_i z_j \Big\} \qquad \text{for all} \quad z \in \mathbb{C}^d,\ x \in \Omega. \tag{3.4.5}$$

The constant c_0 may depend on the coefficient functions $a_{ij}(x)$ as well as d, but not on x or z. Then

$$c_0 \sum_{i=1}^{d} \int_\Omega |\partial_i u|^2\, dx \;\le\; \Re\Big\{ \sum_{i,j=1}^{d} \int_\Omega \partial_i u^* a_{ij}(x) \partial_j u\, dx \Big\} \tag{3.4.6}$$

for all $u \in H^1(\Omega)$.

Proof. Put $z_j = \partial_j u$. For each $x \in \Omega$ we then have

$$c_0 \sum_{i=1}^{d} |\partial_i u|^2 \;\le\; \Re \sum_{i,j=1}^{d} \partial_i \bar{u}\, a_{ij}(x) \partial_j u.$$

We now obtain (3.4.6) by integrating this inequality over Ω. □

3.4.11 Definition. The second order differential operator (3.4.1) is called *elliptic* if its principle part $\sum_{i,j=1}^d \partial_i(a_{ij}\partial_j \cdot)$ satisfies the bound (3.4.5) for some $c_0 > 0$. This same c_0 must work for all $x \in \Omega$ and $z \in \mathbb{C}^d$.

The solution u of a differential equation of the form $Au(x) = f(x)$ remains unchanged if the entire equation is multiplied on the left by a function which is non-zero at each $x \in \Omega$. Many partial differential operators are not elliptic by our definition, but are in fact equivalent to elliptic ones by such a transformation. For instance, the Laplacian Δ is not elliptic, but $-\Delta$ is elliptic.

The following lemma is the key to our first existence theorem for elliptic boundary value problems.

3.4.12 Lemma (upper and lower energy bounds)**.** *Let $\Omega \subset \mathbb{R}^d$ be open and B be the sesqui-linear form in (3.4.3). Assume that the coefficient functions $a_{ij}(x)$, $b_i(x)$, and $c(x)$ of B belong to $L^\infty(\Omega)$. For any $\lambda \in \mathbb{R}$ define the sesqui-linear form $B_\lambda : H^1(\Omega) \times H^1(\Omega) \to \mathbb{C}$ by*

$$B_\lambda(u,v) \;=\; B(u,v) + \lambda \int_\Omega \bar{v} u \, dx \;. \tag{3.4.7}$$

Then B_λ is bounded *on $H^1(\Omega)$: there is a $K > 0$ such that*

$$|B_\lambda(u,v)| \;\leq\; K \|u\|_1 \, \|v\|_1 \tag{3.4.8}$$

for all $u, v \in H^1(\Omega)$.

Assume also the ellipticity condition (3.4.5) holds for $c_0 > 0$. Then there is a $\lambda_0 \in \mathbb{R}$ such that for all $\lambda \geq \lambda_0$, B_λ is coercive *over H^1: there is a $\kappa > 0$ such that*

$$\Re B_\lambda(u,u) \;\geq\; \kappa \|u\|_1^2 \tag{3.4.9}$$

for all $u \in H^1(\Omega)$. In this case

$$\Re B(u,u) \;\geq\; \kappa \, \|u\|_1^2 - \lambda \, \|u\|_0^2 \tag{3.4.10}$$

also holds for all $u \in H^1(\Omega)$.

In this lemma $\lambda_0, K,$ and κ depend on $d, c_0, \|a_{ij}\|_\infty, \|b_i\|_\infty, \|c\|_\infty$, but not on u or v.

Proof. Boundedness. The main tool we need here is Schwarz's inequality

$$|\int_\Omega v^* u \, dx| \leq \big(\int_\Omega |u|^2 \, dx \int_\Omega |v|^2 \, dx\big)^{1/2}$$

which holds for all $u, v \in L^2(\Omega)$. We have

$$\begin{aligned}
|B(u,v)| &\leq \sum_{i=1}^d \sum_{j=1}^d \int_\Omega |\partial_i v| \, |a_{ij}(x)| \, |\partial_j u| \, dx \\
&\qquad + \sum_{i=1}^d \int_\Omega |v| \, |b_i(x)| \, |\partial_i u| \, dx + \int_\Omega |v| \, |c(x)| \, |u| \, dx \\
&\leq \max_{i,j} \big(\sup_{x\in\Omega} |a_{ij}(x)|\big) \sum_{i=1}^d \sum_{j=1}^d \int_\Omega |\partial_i v| \, |\partial_j u| \, dx \\
&\qquad + \max_i \big(\sup_{x\in\Omega} |b_i(x)|\big) \sum_{i=1}^d \int_\Omega |v| \, |\partial_i u| \, dx + \sup_{x\in\Omega} |c(x)| \int_\Omega |v| \, |u| \, dx \\
&\leq C \|u\|_1 \, \|v\|_1 \,,
\end{aligned}$$

where $\|u\|_1^2 = \int_\Omega |u|^2 + |\partial_1 u|^2 + \cdots |\partial_d u|^2 \, dx$ and $C > 0$ depends on r, d, and the sup-norms of the coefficient functions a_{ij}, b_i, and c. Finally,

$$|B_\lambda(u,v)| \leq |B(u,v)| + |\lambda \int_\Omega v^* u \, dx| \leq (C + |\lambda|) \, \|u\|_1 \, \|v\|_1 = K \, \|u\|_1 \, \|v\|_1.$$

Coerciveness. The idea of this part of the proof is the same as saying that if $a > 0$ the polynomial $y = ax^2 + bx + c$ can be made positive for all real x by adding a sufficiently large constant λ to the right hand side: $y = ax^2 + bx + c + \lambda$. In the simple case when all the coefficients $b_i = 0$ the ellipticity assumption on the $a_{ij}(x)$'s means that λ_0 only has to chosen large enough that the function $c(x) + \lambda_0$ is strictly positive on Ω. (This is possible since $c \in L^\infty$.) For the general case λ_0 must be chosen to compensate for the $b_i(x)$'s as well as $c(x)$.

A key inequality needed for this part is based on the familiar $|ab| \le \frac{1}{2}(a^2+b^2)$, true since $(a-b)^2 \ge 0$. If ϵ is *any* positive real number we may set $a = \sqrt{2\epsilon}|u(x)|$, $b = |v(x)|/\sqrt{2\epsilon}$, and integrate to get

$$|\int_\Omega \bar{v}\, u\, dx| \le \epsilon \int_\Omega |u|^2\, dx + \frac{1}{4\epsilon}\int_\Omega |v|^2\, dx \tag{3.4.11}$$

for any $u, v \in L^2(\Omega)$. This amazing inequality is so useful because it allows us to make one of the bounding terms small if we can afford to allow the other to be large.

Now, we denote the real part of a complex number by $\Re$ and use Lemma 3.4.10 and the definition of B to obtain

$$\begin{aligned} c_0 \sum_{i=1}^d \int_\Omega |\partial_i u|^2\, dx &\le \Re\{ \sum_{i,j=1}^d \int_\Omega \partial_i \bar{u}\, a_{ij}(x) \partial_j u\, dx \} \qquad (3.4.12)\\ &= \Re\Big\{ B(u,u) - \sum_{i=1}^d \int_\Omega \bar{u}\, b_i(x) \partial_i u\, dx - \int_\Omega \bar{u}\, c(x) u\, dx \Big\} \\ &\le \Re\{B(u,u)\} + \sum_{i=1}^d \int_\Omega |\bar{u}\, b_i(x) \partial_i u|\, dx + \int_\Omega |\bar{u}\, c(x) u|\, dx \\ &\le \Re\{B(u,u)\} + \sum_{i=1}^d \sup_{x\in\Omega} |b_i(x)| \int_\Omega |u|\, |\partial_i u|\, dx + \sup_{x\in\Omega} |c(x)| \int_\Omega |u|^2\, dx. \qquad (3.4.13) \end{aligned}$$

The last inequality follows as in the above proof that B is bounded.

We next use the 'ϵ-inequality' above to obtain

$$\int_\Omega |u|\, |\partial_i u|\, dx \le \epsilon \int_\Omega |\partial_i u|^2\, dx + \frac{1}{4\epsilon}\int_\Omega |u|^2\, dx. \tag{3.4.14}$$

Substituting (3.4.14) for each i into (3.4.13) gives

$$\begin{aligned} c_0 \sum_{i=1}^d \int_\Omega |\partial_i u|^2\, dx &\le \Re\{B(u,u)\} + \sum_{i=1}^d \epsilon \sup_{x\in\Omega} |b_i(x)| \int_\Omega |\partial_i u|^2\, dx \\ &\quad + \sum_{i=1}^d \frac{1}{4\epsilon} \sup_{x\in\Omega} |b_i(x)| \int_\Omega |u|^2\, dx + \sup_{x\in\Omega} |c(x)| \int_\Omega |u|^2\, dx. \qquad (3.4.15) \end{aligned}$$

Now we want to choose ϵ so small that the second term of (3.4.15) can be subtracted from both sides of (3.4.15) and still not destroy the positiveness of the coefficient in the role of c_0 on the left. To do this choose ϵ so small that

$$\epsilon \sum_{i=1}^d \sup_{x\in\Omega} |b_i(x)| \le c_0/2. \tag{3.4.16}$$

Subtracting the left side of (3.4.16) from both sides of (3.4.15) then implies

$$\begin{aligned} \frac{c_0}{2} \sum_{i=1}^d \int_\Omega |\partial_i u|^2\, dx &\le \Re\{B(u,u)\} + \frac{1}{4\epsilon} \sum_{i=1}^d \sup_{x\in\Omega} |b_i(x)| \int_\Omega |u|^2\, dx + \sup_{x\in\Omega} |c(x)| \int_\Omega |u|^2\, dx \\ &= \Re\{B(u,u)\} + \Big(\frac{1}{4\epsilon} \sum_{i=1}^d \sup_{x\in\Omega} |b_i(x)| + \sup_{x\in\Omega} |c(x)| \Big) \int_\Omega |u|^2\, dx \qquad (3.4.17) \end{aligned}$$

where ϵ is chosen according to (3.4.16). A simple rearrangement of (3.4.17) gives

$$\frac{c_0}{2}\Big(\sum_{i=1}^d \int_\Omega |\partial_i u|^2\, dx + \int_\Omega |u|^2\, dx \Big) - \Big(\frac{1}{4\epsilon} \sum_{i=1}^d \sup_{x\in\Omega} |b_i(x)| + \sup_{x\in\Omega} |c(x)| + \frac{c_0}{2} \Big) \int_\Omega |u|^2\, dx \le \Re\{B(u,u)\}. \tag{3.4.18}$$

So far we have only used B, B_λ has not yet come into play. Finally we add the *real* number $\lambda \int_\Omega |u|^2\,dx$ to both sides of this inequality to get

$$\frac{c_0}{2}\|u\|_1^2 - \Big(\frac{1}{4\epsilon}\sum_{i=1}^d \sup_{x\in\Omega}|b_i(x)| + \sup_{x\in\Omega}|c(x)| + \frac{c_0}{2} - \lambda\Big)\int_\Omega |u|^2\,dx \;\le\; \Re\{B_\lambda(u,u)\}. \tag{3.4.19}$$

Since $|z| \ge \Re z$ for any complex number z this gives the coercive estimate of the Theorem with $\kappa = c_0/2$, as long as we take

$$\lambda \;\ge\; \lambda_0 \;=\; \frac{1}{4\epsilon}\sum_{i=1}^d \sup_{x\in\Omega}|b_i(x)| + \sup_{x\in\Omega}|c(x)| + \frac{c_0}{2}.$$

□

3.4.13 Exercise. Show that $v \mapsto \int_\Omega f\bar v\,dx$ is a bounded conjugate-linear functional on $H_0^1(\Omega)$ when $f \in L^2(\Omega)$.

3.4.14 Exercise. Show that $v \mapsto \int_\Omega f\,\partial_i\bar v\,dx$ is a bounded conjugate-linear functional on $H_0^1(\Omega)$ when $f \in L^2(\Omega)$.

3.4.15 Exercise. Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with smooth boundary, and suppose the boundary $\partial\Omega$ of Ω is partitioned into two non-empty parts, $\partial_1\Omega$ and $\partial_2\Omega$. Let $C^\infty_{(0)}(\overline\Omega)$ denote those $C^\infty(\overline\Omega)$ functions which $= 0$ in a neighborhood of $\partial_1\Omega$. And let V be the completion of $C^\infty_{(0)}(\overline\Omega)$ with respect to the norm of $H^1(\Omega)$. Show that both $C^\infty_{(0)}(\overline\Omega)$ and V are vector spaces. And show that $H_0^1(\Omega) \subset V \subset H^1(\Omega)$.

3.4.16 Exercise. Assume that a sesqui-linear form $B(u,v)$ is coercive over $H^1(\Omega)$, that is, there exists a $\kappa > 0$ such that $\Re B(u,u) \ge \kappa\|u\|_1^2$ for all $u \in H^1(\Omega)$. And let $V \subset H^1(\Omega)$ be a subspace. Show that B is also coercive over V, that is, that there exists a $\kappa > 0$ such that $\Re B(u,u) \ge \kappa\|u\|_1^2$ for all $u \in V$. Assume B is bounded over $H^1(\Omega)$. Show that B is also bounded over V.

3.4.17 Exercise. Let $\Omega \subset \mathbb{R}^d$ be bounded. Show that $L^\infty(\Omega) \subset L^2(\Omega)$ by applying Schwarz's inequality to $\int_\Omega |f(x)g(x)|\,dx$ with $g(x) = 1$ for all $x \in \Omega$.

3.4.18 Exercise. Let $\Omega \subset \mathbb{R}^d$ be open and bounded, and let $z \in \Omega$ and $k \in \{1, 2, ..., d\}$ be fixed. Let $f(x) = 0$ when $x_k < z_k$ and $= 1$ when $x_k \ge z_k$. Interpret $\partial_k f$ and give an explicit formula for $\langle \partial_k f, v\rangle$ when $v \in C_0^\infty(\Omega)$. (You might first consider the case $d = 1$.)

3.4.19 Exercise. Show that Δ is 'formally' negative semi-definite in the sense that $\int \phi\,\Delta\phi\,dx \le 0$ for all $\phi \in C_0^\infty(\Omega)$. (Integrate by parts each term of the form $\int \phi\,\partial_j^2\phi\,dx$ noting that no boundary terms arise because ϕ is zero near the boundary of Ω.)

3.4.3 Solution of the strictly coercive problem

We now come to our first existence theorem. It will play a key role in obtaining further existence theorems later. The assumptions are the same as in Lemma 3.4.12.

3.4.20 Theorem (coercive existence theorem)**.** *Let $\Omega \subset \mathbb{R}^d$ be open and A be given by (3.4.1):*

$$Au = A(x)u(x) = -\sum_{i=1}^d\sum_{j=1}^d \partial_i(a_{ij}(x)\partial_j u) + \sum_{i=1}^d b_i(x)\partial_i u + c(x)u. \tag{3.4.20}$$

Assume that the coefficient functions $a_{ij}(x)$, $b_i(x)$, and $c(x)$ of the differential operator A belong to $L^\infty(\Omega)$ and that A is elliptic (i.e., that (3.4.5) holds). Then there is a $\lambda_0 \in \mathbb{R}$ such that for all $\lambda \ge \lambda_0$ the Dirichlet problem

$$Au + \lambda u = f \tag{3.4.21}$$

has a unique solution $u \in H_0^1(\Omega)$ for any given $f \in H^{-1}(\Omega)$. The λ_0 here is the same as in Lemma 3.4.12.

Proof. Equation (3.4.21) is an equation between two elements of $H^{-1}(\Omega)$. By definition (two functions are equal) this means that

$$\langle Au + \lambda u, v\rangle = \langle f, v\rangle$$

for every $v \in H_0^1(\Omega)$. Lemma 3.4.2 implies that this is equivalent to

$$B_\lambda(u, v) = \langle f, v\rangle$$

for all $v \in H_0^1(\Omega)$. Under the current hypotheses, Lemma 3.4.12 holds and implies that B_λ satisfies the hypotheses of the Lax-Milgram lemma. This shows the existence of a unique u. □

The preceding proof is one of the most important in this manuscript. It is simple because we have developed the right framework for the boundary value problem. We encourage the reader to go through it carefully, checking all details.

3.4.21 Corollary. *Assume the conditions of the preceding theorem. Then for any $\lambda \geq \lambda_0$ the linear operator*

$$A + \lambda I : H_0^1 \to H^{-1} \tag{3.4.22}$$

is a continuous bijection (one-to-one and onto) with continuous inverse.

Proof. Lemma 3.4.1 tells us that the transformation (3.4.22) is continuous. Theorem 3.4.20 shows that the tranformation is one-to-one and onto, so an algebraic inverse

$$(\,A + \lambda I\,)^{-1} : H^{-1} \to H_0^1$$

exists. The open mapping theorem immediately implies that $(A+\lambda I)^{-1}$ is continuous, but a discussion of this point is beyond the scope of these notes. □

In particular, we have

3.4.22 Corollary. *Assume the conditions of the preceding theorem and let $\lambda \geq \lambda_0$. Then for all $f \in L^2(\Omega)$ the equation $Au + \lambda u = f$ has a unique solution $u \in H_0^1(\Omega)$.*

Proof. True since $L^2(\Omega) \subset H^{-1}(\Omega)$. □

In books on PDE it is shown that if $f \in L^2(\Omega)$, the solution of $(A + \lambda)u = f$ in $H_0^1(\Omega)$, actually lies in $H_0^1(\Omega) \cap H^2(\Omega)$. (There are also other solutions of this equation, solutions with non-zero boundary values.)

3.4.23 Example. Let $A = -\Delta + k^2$, where $k^2 > 0$ is a real constant, and let $\Omega \subset \mathbb{R}^d$ be any bounded open set. Then the equation $-\Delta u + k^2 u = f$ has a unique solution $u \in H_0^1(\Omega)$ for any $f \in H^{-1}(\Omega)$, and in particular for any $f \in L^2$.

To prove this we verify the conditions of Theorem 3.4.20. Part (a) of Lemma 3.4.10 is valid, and if $k^2 > 0$ the sesquilinear form

$$B(u, v) \;=\; \int_\Omega \nabla v^* \nabla u + k^2 \bar{v} u \, dx$$

satisfies the coercive estimate

$$B(u, u) \;=\; \int_\Omega |\nabla u|^2 + k^2 |u|^2 \, dx \;\geq\; \kappa \int_\Omega |\nabla u|^2 + |u|^2 \, dx \;=\; \kappa \, \|u\|_1^2$$

if $\kappa = \min\{1, k^2\}$. (Here we use ∇u to denote the column vector of partial derivatives of the complex-valued function u.)

3.4.24 Exercise. Is it true that for any $\lambda > 0$, the operator $\lambda - \Delta$ is coericive over $H_0^1(\Omega)$? Prove or disprove.

3.4.25 Exercise. Generalize the preceding example in the following way. Let

$$A_0 u = -\sum_{i,j=1}^{d} \partial_i (a_{ij}(x) \partial_j u)$$

where $a_{ij} \in L^\infty(\Omega)$ and the matrix $[a_{ij}(x)]$ is uniformly positive definite on $\Omega \subset \mathbb{R}^d$, a bounded open subset. Set $A = A_0 + k^2(x)$ where $k^2 \in L^\infty(\Omega)$ is real and satisfies $k^2(x) \geq \epsilon$ for all $x \in \Omega$. Here, $\epsilon > 0$ is a constant. Show that the equation $Au = f$ has a unique solution $u \in H_0^1(\Omega)$ for any $f \in H^{-1}(\Omega)$, and in particular for any $f \in L^2(\Omega)$.

3.4.26 Exercise. Reach the same solvability conclusion for the equation $Au = f$ in the preceding exercise by using the Riesz representation theorem on $H_0^1(\Omega)$ instead of Lemma 3.4.12 and Theorem 3.4.20.

3.4.27 Exercise. Let (a, b) be a bounded interval in $\mathbb{R}$. Show that

$$\langle u, v \rangle_0' \;=\; \int_a^b u'\, \bar{v}'\, dx$$

is an inner product on $H_0^1(a, b)$ which is equivalent to the usual inner product

$$\langle u, v \rangle_0 \;=\; \int_a^b u\bar{v} + u'\, \bar{v}'\, dx\,.$$

This means to show that there are constants $C, c > 0$ such that

$$C\, \|u\|_0' \;\geq\; \|u\|_0 \;\geq\; c\, \|u\|_0', \qquad \text{for all } u \in H_0^1(a, b)\,.$$

Hint: If $u \in C_0^\infty(a, b)$, $u(x) = \int_a^x u'(t)\, dt$. So for every $x \in (a, b)$, $|u(x)| \leq \int_a^b |u'(t)|\, dt \leq (b-a) \int_a^b |u'|^2\, dt$. (The last inequality is Schwarz's on $L^2(a, b)$.)

3.4.28 Exercise. Generalize the last exercise to $H_0^1(\Omega)$, when $\Omega \subset \mathbb{R}^d$ is bounded and $\langle u, v \rangle_0' = \int_\Omega \nabla u \cdot \nabla \bar{v}\, dx$. (You only need to use a bound like the previous exercise in one of the variables $x_1, x_2, ..., x_{d-1}$, or x_d.)

3.4.29 Exercise. Let $\Omega \subset \mathbb{R}^d$ be bounded and define the Dirichlet form $B(u, v) = \int_\Omega \nabla \bar{v} \cdot \nabla u\, dx$ on $H_0^1(\Omega)$. Use the previous exercise to prove an estimate of the form

$$B(u, u) \;=\; (\, \|u\|_1')^2 \;\geq\; \kappa\, \|u\|_1^2$$

for some $\kappa > 0$. Use the Riesz theorem to show that the equation $-\Delta u = f$ has a unique solution $u \in H_0^1(\Omega)$ for every $f \in H^{-1}(\Omega)$.

3.4.30 Exercise. Generalize the preceding exercise. Let A_0 and Ω be the same as in Exercise 3.4.25 but assume also that Ω is bounded. Show that the equation $A_0 u = f$ has a unique solution $u \in H_0^1(\Omega)$ for any $f \in H^{-1}(\Omega)$.

3.4.31 Exercise. Show that the evaluation linear functional $v \mapsto v(0)$ is not bounded on $L^2(-1, 1)$ but is bounded on $H_0^1(-1, 1)$. (This linear functional is also not bounded on $H_0^1(\Omega)$ when $\Omega \subset \mathbb{R}^d$ contains 0 and $d > 1$.)

3.4.32 Exercise. Consider the biharmonic equation $\Delta^2 u = f$ on a bounded domain Ω. This equation models flexible shells and beams when their displacements are small. In this exercise we will prove that a solution of the biharmonic equation exists in the space $H_0^2(\Omega)$ which, loosely speaking, is the space of $H^2(\Omega)$ functions u which are zero on $\partial\Omega$, and whose first order derivatives $\partial_j u$ are also zero on $\partial\Omega$. For simplicity, assume all functions are real valued.

First, give a definition of $H_0^2(\Omega)$ and $H^{-2}(\Omega)$.

Second, find a Dirichlet form associated with $\Delta^2 = (-\Delta)^2$. Assume $u, v, w \in C_0^\infty(\Omega)$. Use the product rule ($\nabla$ = gradient, $\nabla\cdot$ = divergence)

$$\nabla \cdot (w\, \nabla v) \;=\; w\, \Delta v + \nabla w \cdot \nabla v$$

and the divergence theorem ($\vec{n}$ = outward unit normal)

$$\int_\Omega \nabla \cdot \vec{f}\, dx \;=\; \int_{\partial\Omega} \vec{n} \cdot \vec{f}\, ds \tag{3.4.23}$$

to integrate by parts ($\partial v/\partial n = \vec{n} \cdot \nabla v$)

$$\int_\Omega w\, \Delta v\, dx \;=\; -\int_\Omega \nabla w \cdot \nabla v\, dx + \int_{\partial\Omega} w\, \frac{\partial v}{\partial n}\, ds\,.$$

Now set $v = \Delta u$, and use this product rule again in the form

$$\nabla \cdot [(\nabla w)(\Delta u)] \;=\; \Delta w\, \Delta u + \nabla w \cdot \nabla(\Delta u)$$

to integrate by parts a second time

$$-\int_\Omega \nabla w \cdot \nabla(\Delta u)\, dx \;=\; +\int_\Omega (\Delta w)(\Delta u)\, dx - \int_{\partial\Omega} (\frac{\partial w}{\partial n})(\Delta u)\, ds\,.$$

Since w and $\partial w/\partial n$ are zero on the boundary, the integrals over $\partial\Omega$ go away and we are left with

$$\int_\Omega w\, \Delta^2 u\, dx \;=\; \int_\Omega (\Delta w)(\Delta u)\, dx\,, \qquad \text{all } u, w \in C_0^\infty(\Omega)\,. \tag{3.4.24}$$

Third, show that

$$\langle u, w\rangle_2 \;=\; \int_\Omega (\Delta w)(\Delta u)\, dx$$

is an inner product on $C_0^\infty(\Omega)$. You need to use the following fact: if $u \in H^2(\Omega)$, $\Delta u = 0$ on Ω, and $u = 0$ on $\partial\Omega$, then $u = 0$ on Ω. (See the sentence following Corollary 3.4.22 and Exercise 3.4.29.)

Fourth, show that (3.4.24) 'holds' for all u, w in $H_0^2(\Omega)$. Use the fact that $C_0^\infty(\Omega)$ is dense in $H_0^2(\Omega)$ (by definition of $H_0^2(\Omega)$), and both sides of (3.4.24) are continuous. Note, the left side of (3.4.24) makes no sense when u and w are $H_0^2(\Omega)$ functions; it needs to be replaced by a duality pairing where the second Δ in $\Delta^2 u$ is a weak derivative, and is *defined* by the right side.

Fifth, use the Riesz representation theorom to show that $\Delta^2 u = f$ has a weak solution u in $H_0^2(\Omega)$. Assume $f \in L^2(\Omega)$ and show that it is a bounded linear finctional on $H_0^2(\Omega)$. What boundary conditions does u satisfy?

3.4.33 Exercise (solid mechanics and electromagnetics)**.** The static deformation of elastic solids (assuming small displacements) which are homogeneous and isotropic, or the static displacement of electromagnetic vector fields in homogeneous and isotropic media, satisfy an equation of the form

$$\underset{d\times 1}{A\, u} = -\alpha\Delta u - \beta\nabla\, \nabla \cdot u \;=\; \underset{d\times 1}{f(x)}.$$

Here α and β are positive constants and $d = 1, 2$ or 3. Define an inner product on $H_0^1(\Omega; \mathbb{C}^d)$ that is a Dirichlet form for this PDE. Speculate on the solvability of this equation.

3.4.4 Complete solution of the Dirichlet problem; eigen-values

In this section we generalized the solvability results of the preceding section.

If X and Y are vector spaces and, as sets, $X \subset Y$, then there is a natural one-to-one linear transformation, $E : X \to Y$, that *embeds* X in Y. That is, $E(X) = X \subset Y$ (or $\mathscr{R}(E) = X \subset Y$) and $Eu = u \in Y$ for all $u \in X$. We will say that X is *embedded* in Y and that E is an *embedding* of X into Y. The explicit inclusion of an embedding operator E in an equation is usually omitted, but sometimes clarifies the operations involved.

3.4.34 Lemma. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, A satisfy the conditions of Theorem 3.4.20, and $E : H_0^1(\Omega) \to H^{-1}(\Omega)$ denote the embedding. If $\lambda \in \mathbb{R}$ is so large that $(A + \lambda E) : H_0^1(\Omega) \to H^{-1}(\Omega)$ is a bijection then the composition*

$$(A + \lambda E)^{-1} \circ E : H_0^1(\Omega) \to H_0^1(\Omega)$$

is compact.

Proof. Consider the following diagram

$$H_0^1 \xrightarrow{E_1} L^2 \xrightarrow{E_2} H^{-1} \xrightarrow{(A+\lambda E)^{-1}} H_0^1$$

which illustrates the composition of three bounded linear transformations. All are continuous, and E_1 is compact by the Rellich embedding theorem. Since the composition of a bounded mapping and a compact one is compact, the composition $E = E_2 \circ E_1 : H_0^1 \to H^{-1}$ is compact, and so is $(A + \lambda E)^{-1} \circ E$. □

The *formal adjoint* A^* of the differential operator A is defined by the equation

$$\int_\Omega \bar{v} A u \, dx \;=\; \int_\Omega (\overline{A^* v}) u \, dx$$

for all $u, v \in C_0^\infty(\Omega)$. Integration by parts shows that

$$A^* v \;=\; -\sum_{i,j=1}^{d} \partial_i(\bar{a}_{ji}\partial_j v) - \sum_{i=1}^{d} \partial_i(\bar{b}_i v) + \bar{c} v \,. \tag{3.4.25}$$

Notice that the matrix $[\bar{a}_{ji}]$ of the leading term of A^* is the conjugate transpose of that of A.

If A is elliptic, i.e., satisfies (3.4.5), then A^* is also elliptic (with perhaps a different c_0). (Exercise 3.4.40.)

Define the null space (or kernel) of A and A^* as

$$\mathscr{N}(A) \;=\; \{u \in H_0^1(\Omega) \,;\; Au = 0\},$$

$$\mathscr{N}(A^*) \;=\; \{u \in H_0^1(\Omega) \,;\; A^* u = 0\}.$$

The dimensions of these vector spaces will be denoted $\dim \mathscr{N}(A)$ and $\dim \mathscr{N}(A^*)$.

Let X and Y be vector spaces with X embedded in Y, and let $A : X \to Y$ be a linear transformation. We will say that $\lambda \in \mathbb{C}$ is an *eigen-value* of A if there is a $w \in X$ such that $w \neq 0$ and $Aw = \lambda\, Ew$ where E is the embedding of X into Y. Such a w will be called an *eigen-vector* or *eigen-function.*

We now use the Fredholm alternative to obtain more information about A.

3.4.35 Theorem (complete existence theorem)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded open subset and A be given by (3.4.1) or (3.4.20). Assume that the coefficient functions $a_{ij}(x)$, $b_i(x)$, and $c(x)$ of the differential operator A belong to $L^\infty(\Omega)$, and that A is elliptic, i.e., satisfies (3.4.5). Then the bounded linear operator $A : H_0^1(\Omega) \to H^{-1}(\Omega)$ has a countable set $\{\lambda_k \,;\; k = 1, 2, ...\} \subset \mathbb{C}$ of eigenvalues. For each eigenvalue λ_k there exist non-zero functions $w_k \in H_0^1(\Omega)$ satisfying*

$$Aw_k \;=\; \lambda_k w_k. \tag{3.4.26}$$

The set $\{\lambda_k\}_{k=1}^\infty$ satisfies $\operatorname{Re}\lambda_k \to \infty$ as $k \to \infty$.

Moreover, for any complex number $\lambda \notin \{\lambda_k \,;\; k \in \mathbb{N}\}$, the equation

$$Au - \lambda u \;=\; f \tag{3.4.27}$$

has a unique solution $u \in H_0^1(\Omega)$ for any $f \in H^{-1}(\Omega)$. And if λ_k is an eigen-value of A, the equation

$$Au - \lambda_k u \;=\; f \tag{3.4.28}$$

has a (non-unique) solution $u \in H_0^1(\Omega)$ if and only if $\langle f, v\rangle = 0$ for every $v \in H_0^1(\Omega)$ satisfying $A^* v - \overline{\lambda}_k v = 0$.

Proof. Throughout the proof we keep in mind the following embeddings:

$$H_0^1(\Omega) \xrightarrow{E_1} L^2(\Omega) \xrightarrow{E_0} H^{-1}(\Omega)$$

$$H_0^1(\Omega) \xrightarrow{E = E_0\, E_1} H^{-1}(\Omega).$$

We know that E_1 is compact by Rellich's theorem (introduction to Section 3.4). And each embedding is one-to-one so that $Eu = Ev$ in $H^{-1}(\Omega)$, for instance, implies $u = v$ in $H_0^1(\Omega)$.

By Theorem 3.4.20 and Exercise TBD we may take $\lambda_0 \in \mathbb{R}$ so large that both

$$H_0^1(\Omega) \xrightarrow{A+\lambda_0 E} H^{-1}(\Omega).$$

$$H_0^1(\Omega) \xrightarrow{A^*+\lambda_0 E} H^{-1}(\Omega).$$

are continuous, one-to-one, and onto, with continuous inverses. And we will need to use the fact that $[(A + \lambda E)^{-1}]^* = (A^* + \overline{\lambda}E)^{-1}$ but we will omit the tedious argument.

Now we write the equation $(A - \lambda)u = f$, in the vector space $H^{-1}(\Omega)$, as

$$(\, A + \lambda_0 E\,)u - (\lambda + \lambda_0)Eu \;=\; f\,.$$

Applying $(\, A + \lambda_0 E\,)^{-1}$ on the left we obtain

$$u - (\lambda + \lambda_0)(\, A + \lambda_0 E\,)^{-1} E\, u \;=\; (\, A + \lambda_0 E\,)^{-1}\, f\,, \tag{3.4.29}$$

an equation in $H_0^1(\Omega)$. Thus, $u \in H_0^1(\Omega)$ solves the equation $Au = f$ if and only if u solves the equation (3.4.29). And finally if we apply the compact operator E_1 to this equation we have

$$E_1 u - (\lambda + \lambda_0)E_1(\, A + \lambda_0 E\,)^{-1} E_0 E_1\, u \;=\; E_1(\, A + \lambda_0 E\,)^{-1}\, f\,, \tag{3.4.30}$$

an equation in $L^2(\Omega)$.

Now the operator $K = E_1(\, A + \lambda_0 E\,)^{-1} E_0$ is a compact transformation of $L^2(\Omega)$ into itself because it is the composition of a bounded transformation and a compact one. And (3.4.30) can be written

$$[I - (\lambda + \lambda_0)\, K]\, v \;=\; g \tag{3.4.31}$$

where $v = E_1 u$ is nothing but the function u in $L^2(\Omega)$ and $g = E_1(\, A+\lambda_0 E\,)^{-1}\, f$. Finally, setting $\mu = (\lambda+\lambda_0)^{-1}$. so long as $\lambda \neq \lambda_0$, we may write (3.4.31) in the form

$$(\mu - K)\, v \;=\; \mu\, g \tag{3.4.32}$$

to which the Fredholm theorem applies.

In order to apply Fredholm's theorem we play the same game to change the form of the adjoint equation $(A^* - \overline{\lambda})u = f$. We are lead to the equivalent equation

$$E_1 u - (\overline{\lambda} + \lambda_0)E_1(\, A^* + \lambda_0 E\,)^{-1} E_0 E_1\, u \;=\; E_1(\, A^* + \lambda_0 E\,)^{-1}\, f\,, \tag{3.4.33}$$

in $L^2(\Omega)$ (remember λ_0 is real), or more compactly

$$(\overline{\mu} - K^*)\, w \;=\; \overline{\mu}\, h \tag{3.4.34}$$

where $K^* = E_1(\, A^* + \lambda_0 E\,)^{-1} E_0$ is the adjoint of K although we do not verify this fact.

We can now apply Fredholm's theorem; the right hand sides of (3.4.32) and (3.4.34) play no role in the theory. We conclude the following:

(a) The null space $\mathscr{N}(\mu - K)$ is $0 \in L^2(\Omega)$ except for a finite or countably infinite set $\{\mu_k\ ;\ k \in \mathbb{N}\}$. And, if infinite, $\mu_k \to 0$ as $k \to \infty$. Further, the dimension $\dim\mathscr{N}(\mu_k - K) < \infty$ for all $k \in \mathbb{N}$.

(b) If $\mu \neq 0$, $\dim\mathscr{N}(\mu - K) = \dim\mathscr{N}(\overline{\mu} - K^*)$.

(c) If $\mu \neq 0$, both range spaces $\mathscr{R}(\mu - K)$ and $\mathscr{R}(\overline{\mu} - K^*)$ are closed in $L^2(\Omega)$.

(d) If $\mu \neq 0$, the equation $(\mu - K)v = g$ has a solution if and only if $g \perp \mathscr{N}(\overline{\mu} - K^*)$ in $L^2(\Omega)$, that is, $\int_\Omega g(x)\overline{w}(x)\, dx = 0$ for every w which satisfies $(\overline{\mu} - K^*)w = 0$. And, $(\overline{\mu} - K^*)w = h$ has a solution w if and only if $h \perp \mathscr{N}(\mu - K)$ in $L^2(\Omega)$.

(e) If $\mu \neq 0$, the following are equivalent: (i) the operator $(\mu - K) : L^2(\Omega) \to L^2(\Omega)$ is onto; (ii) the operator $(\overline{\mu} - K^*) : L^2(\Omega) \to L^2(\Omega)$ is onto; (iii) the operator $(\mu - K) : L^2(\Omega) \to L^2(\Omega)$ is one-to-one; (iv) the operator $(\overline{\mu} - K^*) : L^2(\Omega) \to L^2(\Omega)$ is one-to-one; (v) $\mu \notin \{\mu_k\ ;\ k \in \mathbb{N}\}$, the set of eigen-values of K.

Now we interpret these points for the differential equation (3.4.27).

(a') The differential operator A has at most a countable number of eigen-values $\lambda_k = 1/\mu_k - \lambda_0$, $k \in \mathbb{N}$, and $\lambda_k \to \infty$ in $\mathbb{C}$ as $k \to \infty$. Each eigen-space of A is finite dimensional. (We may therefore choose an orthonormal set of $L^2(\Omega)$ functions as a basis for it.)

(b') Since $\frac{1}{\lambda - \lambda_0} \neq 0$ is always true, the eigen-spaces of A and A^* which correspond to eigen-values λ_k and $\overline{\lambda}_k$ have the same dimension.

(c') for all $\lambda \in \mathbb{C}$, the range spaces of $A - \lambda$ and $A^* - \overline{\lambda}$, or just $A^* - \lambda$, are both closed in $H^{-1}(\Omega)$.

(d') The equation (3.4.27), $(A - \lambda)u = f$, has a solution if and only if $f \perp \mathscr{N}(A^* - \overline{\lambda})$ (in the sense of the duality pairing).

(e') The following are equivalent: (i') $(A - \lambda) : H_0^1(\Omega) \to H^{-1}(\Omega)$ is onto; (ii') $(A^* - \overline{\lambda}) : H_0^1(\Omega) \to H^{-1}(\Omega)$ is onto; (iii') $(A - \lambda) : H_0^1(\Omega) \to H^{-1}(\Omega)$ is one-to-one; (iv') $(A^* - \overline{\lambda}) : H_0^1(\Omega) \to H^{-1}(\Omega)$ is one-to-one; (v') $\lambda \notin \{\lambda_k \, ; \, k \in \mathbb{N}\}$, the set of eigen-values of A.

As a final point, we observe that each eigen-function of A, satisfying $(A - \lambda_0)w_k = (\lambda_k - \lambda_0)w_k$, and which from the analysis of K has only been shown to be in $L^2(\Omega)$, is actually in $H_0^1(\Omega)$. For the right side $(\lambda_k - \lambda_0)w_k$ of this eigen-equation is in $L^2(\Omega)$ hence in $H^{-1}(\Omega)$, and $(A - \lambda_0)$ is a bijection, so the solution w_k must lie in $H_0^1(\Omega)$.

This verifies all claims made in the theorem. □

Diffraction Problems If a wave is propagating in a medium, that wave can be diffracted along an internal hypersurface where the medium changes properties. In a differential equation model, the coefficients of the equation, e.g., the speed of wave propagation, change value across this surface. The theory for solving elliptic boundary value problems that we have laid down applies when the coefficients of the Dirichlet form are merely L^∞ functions; they need not be continuous for the Lax-Milgram lemma to apply.

Suppose for instance that we want to solve the equation for standing waves, the Helmholts equation

$$\Delta u + \frac{\omega^2}{c^2} u \;=\; f(x)\,,$$

in a region $\Omega \subset \mathbb{R}^d$, where ω is the wave frequency and c is the speed of wave propagation in the medium. Let $\Omega = \Omega_1 \cup \Omega_2$, and the wave speeds be c_1 and c_2 in Ω_1 and Ω_2. Define wave numbers $k_i = \omega/c_i$ for $i = 1, 2$, and the wave number function in Ω by the step function $k(x) = k_1\, 1_{\Omega_1}(x) + k_2\, 1_{\Omega_2}(x)$ where 1_{Ω_i} is the indicator function for the set Ω_i. We rewrite the differential equation

$$(I - \Delta)u - (1 + k^2(x))u \;=\; -f \qquad \text{or} \qquad u - Ku \;=\; v$$

where $K = (I - \Delta)^{-1}(1 + k^2(x))$ is compact and $v = -(I - \Delta)^{-1} f$.

From here, we conclude that K has a discrete set of (possibly complex) eigen-values λ_j converging to zero. If $\lambda_j \neq 1$ for any j the Helmholtz equation has a unique solution $u \in H_0^1(\Omega)$ for any $f \in H^{-1}(\Omega)$ (for instance).

3.4.36 Exercise. Prove formula (3.4.25) for A^*. When integrating by parts, remember that u and v vanish in some neighborhood of $\partial\Omega$, so no 'boundary evaluations' are needed.

3.4.37 Exercise. Find conditions on $a_{ij}(x)$, $b_i(x)$, and $c(x)$ such that $A^* = A$.

3.4.38 Exercise. If $A^* = A$, show that $(u, v) \mapsto \int_\Omega \bar{v}\, Au\, dx$ is an inner product on $C_0^\infty(\Omega)$, assuming also that the matrix $[a_{ij}(x)]$ is positive definite and the function $c(x)$ is positive.

3.4.39 Exercise. Under the conditions of the previous exercise, show that the inner product $\int_\Omega \bar{v}\, Au\, dx$ is equivalent to the inner product $(u, v) \mapsto \int_\Omega \bar{v}u + \sum_i \partial_i \bar{v}\, \partial_i u\, dx$ on $C_0^\infty(\Omega)$.

3.4.40 Exercise. If A given by (3.4.1) is elliptic (satisfies (3.4.5)), show that A^* given by (3.4.25) is also elliptic. (Hint: use the smallest singular value of the matrix $[a_{ij}(x)]$.)

3.4.41 Example (Helmholtz equation)**.** Let $A = -\Delta - k^2$ where $k^2 > 0$ is a real constant, and let $\Omega \subset \mathbb{R}^d$ be any bounded open set. We cannot conclude from Theorem 3.4.20 that the equation $Au = f$ is solvable for $u \in H_0^1(\Omega)$ when given any $f \in H^{-1}(\Omega)$. The best we can do is to conclude that the equation $Au + \lambda u = f$ is solvable for any $\lambda \geq k^2$; this is the conclusion of Examples 3.4.23 and 3.4.29. We will be able to conclude more from our results in the next section.

3.4.42 Exercise (Helmholtz equation with variable wave speed)**.** Let Ω be an open, bounded set, and $A = -\Delta - k(x)^2$ where the function $k^2 \in L^\infty(\Omega)$ satisfies $0 \le k(x)^2 \le c$ for all $x \in \Omega$ with some constant $c > 0$. Find the largest interval $(a, b) \subset \mathbb{R}$ that you can which has the property that the equation $Au + \lambda u = f$ is uniquely solvable for all $f \in H^{-1}(\Omega)$ whenever $\lambda \in (a, b)$. (Either a or b may be infinite.)

3.4.43 Exercise. Repeat the preceding exercise with the operator $-\Delta$ replaced by A_0 in Exercise 3.4.25.

3.4.5 Non-homogeneous Dirichlet boundary conditions

In this section we show how to reduce the non-homogeneous Dirichlet problem

$$Au = f \qquad \text{on } \Omega \tag{3.4.1}$$
$$u = g \qquad \text{on } \partial\Omega, \tag{3.4.2}$$

where g is a given function on $\partial\Omega$, to the homogeneous problem

$$Av = \tilde{f} \qquad \text{on } \Omega \tag{3.4.3}$$
$$v = 0 \qquad \text{on } \partial\Omega, \tag{3.4.4}$$

where $\tilde{f} = f - A\tilde{g}$ and $\tilde{g}$ is any function in $H^1(\Omega)$ such that $\tilde{g} = g$ on $\partial\Omega$. If we can solve for v, the function $u = v + \tilde{g}$ then solves the non-homogeneous boundary problem. We say that $u = \tilde{g}$ on $\partial\Omega$ when u and $\tilde{g}$ belong to $H^1(\Omega)$ and $u - \tilde{g} \in H^1_0(\Omega)$. And we call the function $\tilde{g}$ the 'generalized boundary values' of u on $\partial\Omega$.

3.4.44 Theorem. *Let A be an elliptic differential operator on a bounded domain Ω in $\mathbb{R}^d$ satisfying the hypotheses of Theorem* **??**. *And let us denote the restriction of the bounded linear transformation $A : H^1(\Omega) \to H^{-1}(\Omega)$ to the subspace $H^1_0(\Omega)$ by $A_0 : H^1_0(\Omega) \to H^{-1}(\Omega)$ (cf. Lemma 3.4.1). (A and A_0 are the* same *differential operators here, both given by (3.4.1) or (3.4.20).) Assume that 0 is not an eigenvalue of A_0 (cf. Theorem* **??***), and let $g \in H^1(\Omega)$ and $f \in H^{-1}(\Omega)$ be given. Then the (non-homogeneous) Dirichlet problem: find $u \in H^1(\Omega)$ which satisfies*

$$Au = f \qquad \text{and} \qquad u - g \in H^1_0(\Omega), \tag{3.4.5}$$

has a unique solution. And it is given by $u = v + g$ where $v \in H^1_0(\Omega)$ is the unique solution to

$$A_0 v = \tilde{f} \tag{3.4.6}$$

with $\tilde{f} = f - Ag \in H^{-1}(\Omega)$.

Proof. By Lemma 3.4.1, $A : H^1 \to H^{-1}$ is continuous, and if $g \in H^1$ then $Ag \in H^{-1}$. So $\tilde{f} \in H^{-1}$ if f is. By Theorem **??**, if 0 is not an eigenvalue of A_0 the homogeneous Dirichlet problem, $A_0 v = \tilde{f}$, has a unique solution (which of course depends on g).

Now with v in hand, set $u = v + g$; this is an element of H^1. Clearly $u - g = v \in H^1_0$ so that this u satisfies the correct Dirichlet boundary conditions. And we may compute

$$Au \;=\; A(v+g) \;=\; A_0 v + Ag \;=\; (f - Ag) + Ag \;=\; f$$

so that u satisfies the elliptic differential equation in Ω.

Now we address the uniqueness of the solution u. Suppose g_1 and g_2 both belong to H^1 and have the same boundary values in the sense that $g_1 - g_2 \in H^1_0$. If v_1 and v_2 are the corresponding solutions to (3.4.6), is it true that $v_1 + g_1 = v_2 + g_2$? That is, is the solution u to (3.4.5) well defined in the sense that it is independent of the g used to calculate v (so long as g has the correct boundary values)?

To see that the answer is 'yes' observe that $v_1 + g_1 = v_2 + g_2$ if and only if $v_1 - v_2 + g_1 - g_2 = 0$ as an equation in the vector space $H^1(\Omega)$. But the left side of this equation lies in the subspace $H^1_0(\Omega)$ since both $v_1 - v_2$ and $g_1 - g_2$ do. Therefore

$$A_0(v_1 - v_2 + g_1 - g_2) \;=\; A_0(0) \;=\; 0$$

which shows that indeed $v_1 - v_2 + g_1 - g_2 = 0$, since A_0 is one-to-one. □

3.4.6 Neumann boundary conditions

The Neumann boundary value problem for the Laplacian is to find the function u on Ω which satisfies

$$-\Delta u = f \qquad \text{in } \Omega \tag{3.4.7}$$

$$\frac{\partial u}{\partial n} = g \qquad \text{on } \partial\Omega, \tag{3.4.8}$$

where g is a given function on $\partial\Omega$ and n is the unit outward normal to $\partial\Omega$. As we will see shortly there is a certain naturalness to this boundary condition. However, for more general elliptic operators the normal n needs to be replaced by the 'co-normal' $\tilde{n}$. The components of $\tilde{n}$ involve the second order coefficients a_{ij} of A; $\tilde{n}_j = \sum_{i=1}^d n_i a_{ij}$. To see what motivates this relation we need to extend Lemma 3.4.2 to the case that u and v do not have compact support.

3.4.45 Lemma. *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with piecewise C^1 boundary $\partial\Omega$. Let A be given by (3.4.1) and have coefficient functions $a_{ij}(x) \in C^1(\bar{\Omega})$, and $b_i(x), c(x) \in L^\infty(\Omega)$. And let the Dirichlet sesqui-linear form B be given by (3.4.3). Then the equation*

$$\int_\Omega v^* Au\, dx \;=\; B(u,v) - \int_{\partial\Omega} v^* \frac{\partial u}{\partial \tilde{n}}\, ds \tag{3.4.9}$$

holds for all $u, v \in C^2(\bar{\Omega})$, where the 'co-normal' derivative is

$$\frac{\partial}{\partial \tilde{n}} \;=\; \sum_{i,j=1}^d n_i a_{ij} \partial_j. \tag{3.4.10}$$

Proof. The proof is the same as the proof of Lemma 3.4.2 except that the integration by parts includes boundary terms. If we let $\vec{f} = f(x)\vec{e}_i$ in the divergence theorem (see the introduction to Section 3.4). we have

$$\int_\Omega \partial_i f(x)\, dx \;=\; \int_{\partial\Omega} n_i f(x)\, ds(x)\,.$$

To make the integration by parts clear, let's write out the product rule in the case of interest to us. For each $u \in C^2(\overline{\Omega})$, $v \in C^1(\overline{\Omega})$, and $a_{ij} \in C^1(\overline{\Omega})$ for $i, j = 1, ..., d$, we have

$$\partial_i\big(v^* a_{ij} \partial_j u\big) = (\partial_i v^*) a_{ij} \partial_j u + v^* \partial_i\big(a_{ij} \partial_j u\big).$$

Thus, for each fixed i and j,

$$\begin{aligned} -\int_\Omega v^* \partial_i (a_{ij} \partial_j u)\, dx \;&=\; \int_\Omega (\partial_i v^*) a_{ij} \partial_j u\, dx - \int_\Omega \partial_i (v^* a_{ij} \partial_j u)\, dx \\ &=\; \int_\Omega \partial_i v^* a_{ij} \partial_j u\, dx - \int_{\partial\Omega} n_i (v^* a_{ij} \partial_j u)\, ds(x). \end{aligned}$$

Summing over both i and j gives

$$-\int_\Omega v^* [\sum_{i,j=1}^d \partial_i (a_{ij} \partial_j u)]\, dx \;=\; \sum_{i,j=1}^d \int_\Omega \partial_i v^* a_{ij} \partial_j u\, dx - \int_{\partial\Omega} v^* [\sum_{i,j=1}^d n_i a_{ij} \partial_j u]\, ds(x)$$

Adding the lower order terms of A, which require no partial integration, gives (3.4.9). □

In applications the domain Ω often has corners. So that we may state the next proposition in greater generality we introduce the following notation. Let Ω be a bounded domain with piecewise C^1 boundary, and define the smooth boundary $\partial\tilde{\Omega}$ of Ω to be those points of $\partial\Omega$ which lie in a C^1 piece of $\partial\Omega$. That is, $\partial\tilde{\Omega}$ is the set of all points of $\partial\Omega$ except those that lie on 'corners.' Also, define the smooth closure $\tilde{\Omega}$ to be the union of Ω and its smooth boundary, $\Omega \cup \partial\tilde{\Omega}$.

3.4.46 Proposition. *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with piecewise C^1 boundary $\partial\Omega$. Let A be given by (3.4.1) and have coefficient functions $a_{ij}(x) \in C^1(\bar{\Omega})$, and $b_i(x), c(x) \in C(\bar{\Omega})$. Let the Dirichlet sesqui-linear form B be given by (3.4.3) and $\partial/\partial\tilde{n}$ the co-normal derivative defined in Lemma 3.4.45. Finally, let $u \in C^2(\tilde{\Omega})$, $f \in C(\tilde{\Omega})$, and $g \in C(\partial\tilde{\Omega})$. Then u satisfies*

$$Au \;=\; f \quad \text{on } \Omega \qquad\qquad \text{and} \qquad\qquad \frac{\partial u}{\partial \tilde{n}} \;=\; g \quad \text{on } \partial\tilde{\Omega} \tag{3.4.11}$$

if and only if

$$B(u,v) \;=\; \int_\Omega v^* f\, dx + \int_{\partial\Omega} v^* g\, ds(x) \tag{3.4.12}$$

holds for all $v \in C^\infty(\overline{\Omega})$.

The equations in (3.4.11) are equations in the vector space of continuous functions on Ω and $\partial\tilde{\Omega}$, respectively. The preceding proposition *does not imply* that, given Ω, f, and g satisfying the hypotheses, there then exists a $u \in C^2$ which satisfies the boundary value problem (3.4.11).

Proof. By Lemma 3.4.45, (3.4.12) is equivalent to

$$\int_\Omega v^* A(x) u\, dx + \int_{\partial\Omega} v^* \frac{\partial u}{\partial \tilde{n}}\, ds(x) \;=\; \int_\Omega v^* f\, dx + \int_{\partial\Omega} v^* g\, ds(x). \tag{3.4.13}$$

The *only if* part is now obvious since the portion of $\partial\Omega$ which is not in $\partial\tilde{\Omega}$ has measure zero.

The proof of the *if* part makes use of the 'fundamental lemma of the calculus of variations' which says that if f and g are two continuous functions on a set $S \subset \mathbb{R}^d$ then the relation

$$\int_S \phi\, f\, ds \;=\; \int_S \phi\, g\, ds$$

for all $\phi \in C^\infty(S)$ implies that $f = g$ on S. Here, ds is the volume or surface area element on S. Assuming this, let $v \in C_0^\infty(\Omega) \subset C^\infty(\bar{\Omega})$. By (3.4.13),

$$\int_\Omega v^* Au\, dx \;=\; \int_\Omega v^* f\, dx\,.$$

Since v is arbitrary and Au and f are continuous, $Au = f$ on Ω.

We have just shown that the first terms on each side of equation (3.4.13) are always equal to each other; they may therefore be cancelled to give

$$\int_{\partial\Omega} v^* \frac{\partial u}{\partial \tilde{n}}\, ds \;=\; \int_{\partial\Omega} v^* g\, ds \tag{3.4.14}$$

for all $v \in C^\infty(\bar{\Omega})$. Now on any portion of $\partial\Omega$ that is C^1, every tangent vector to $\partial\Omega$, and therefore the normal n, is continuous. Under our assumptions on the a_{ij} and on u, the function $\partial u/\partial\tilde{n}$, defined by (3.4.10), is therefore continuous on these portions as well. If the two (in general, complex, vector-valued) functions $\partial u/\partial\tilde{n}$ and g are not equal on $\partial\tilde{\Omega}$, they must differ in the real or imaginary part at some $x_0 \in \partial\tilde{\Omega}$. By continuity, $\Re\,(\partial u/\partial\tilde{n} - g) \geq \epsilon > 0$, for instance, on some δ-neighborhood, $\{x \in \mathbb{R}^d\,;\; |x - x_0| < \delta\} \cap \partial\tilde{\Omega}$ in $\partial\tilde{\Omega}$.

Now take a real valued function $v \in C_0^\infty(\mathbb{R}^d)$ with support in $\{|x - x_0| < \delta\}$, and satisfying $0 \leq v(x) \leq 1$ for all $x \in \mathbb{R}^d$ and $v \equiv 1$ on $\{|x - x_0| < \delta/2\}$. By its construction we see that

$$\Re \int_{\partial\Omega} v\big(\frac{\partial u}{\partial \tilde{n}} - g\big)\, ds \;\geq\; \epsilon\, \mathrm{vol}(\{|x - x_0| < \delta/2\}) \;>\; 0$$

contradicting (3.4.14). □

Classically the *Neumann boundary value problem* is to find u satisfying (3.4.11). Except for the technical difficulty of not knowing ahead of time that $u \in C^2$, the preceding proposition shows us exactly how to use the Dirichlet form $B(\ ,\ )$ to solve the Neumann problem. In general, even when the smoothness of u is not known a priori, we shall make the following definition.

3.4.47 Definition. We say that a function u *satisfies the generalized Neumann boundary value problem* (3.4.11) if u satisfies (3.4.12) for every $v \in C^\infty(\bar{\Omega})$.

Before proving our solvability theorems for the generalized Neumann problem we need a lemma.

3.4.48 Lemma. *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with C^1 boundary, and let $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$. Then the conjugate-linear mapping*

$$v \mapsto \int_\Omega v^* f\, dx + \int_{\partial\Omega} v^* g\, ds \tag{3.4.15}$$

defines a bounded functional on $H^1(\Omega)$.

Proof. We have already shown that $v \mapsto \int_\Omega f\,\overline{v}\, dx$ is bounded in the course of proving solutions of the Dirichlet problem. We have to check the boundedness of the second term.

We must first show that there is a $c > 0$ such that

$$\|\overline{v}\|_{L^2(\partial\Omega)} \;\le\; c\,\|\overline{v}\|_{H^1(\Omega)} \tag{3.4.16}$$

for all $v \in H^1(\Omega)$. Because the surface $\partial\Omega$ is curved we first decompose it using a partition of unity, then straighten the pieces by continuously differentiable transformations which map the pieces of $\partial\Omega$ into $S = \{x \in \mathbb{R}^d \,;\, x_d = 0, |x'| < r\}$ where $x' = (x_1, ..., x_{d-1})$, $x = (x', x_d)$, and $r > 0$ is fixed. Without displaying all these (tedious) details, this reduces the problem to showing

$$\int_S |v(x',0)|^2\, dx' \;\le\; C \int_{\{x_d>0\}} |v(x)|^2 + \sum_1^d |\partial_j v(x)|^2\, dx \tag{3.4.17}$$

for some $C > 0$, whenever the support of v lies in the ball $|x| < r$ in $\mathbb{R}^d$. To show this, use the fundumental theorem of calculus: $v(x',0) \;=\; -\int_0^\infty \partial_d v(x',t)\, dt$. Then

$$|v(x',0)|^2 \;\le\; C' \int_0^\infty |\partial_d v(x',t)|^2\, dt$$

for some C' which depends on r but not v. Integrating this inequality over S gives (3.4.17).

Finally using (3.4.16) we have

$$\Big|\int_{\partial\Omega} g\overline{v}\, ds\Big| \le \Big(\int_{\partial\Omega} |g|^2\, ds\Big)^{1/2} \Big(\int_{\partial\Omega} |\overline{v}|^2\, ds\Big)^{1/2} = \|g\|_{L^2(\partial\Omega)}\, \|v\|_{L^2(\partial\Omega)} \;\le\; c\,\|g\|_{L^2(\partial\Omega)}\, \|v\|_{H^1(\Omega)}$$

so the map $v \mapsto \int_{\partial\Omega} g\,\overline{v}\, ds$ is bounded. □

3.4.49 Theorem (coercive existence theorem)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with piecewise smooth boundary, and let $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$. Then there is a $\lambda_0 \ge 0$ such that the generalized Neumann problem, find $u \in H^1(\Omega)$ which satisfies*

$$B_\lambda(u,v) \;\equiv\; B(u,v) + \lambda \int_\Omega v^* u\, dx \;=\; \int_\Omega v^* f\, dx + \int_{\partial\Omega} v^* g\, ds \tag{3.4.18}$$

for every $v \in C^\infty(\overline{\Omega})$, has a unique solution whenever $\lambda \ge \lambda_0$.

Proof. Since B is coercive over $H^1(\Omega)$ there is a $\lambda_0 \ge 0$ such that B_λ is strictly coercive over H^1 for all $\lambda \ge \lambda_0$. Now combine Lemma 3.4.48 with the Lax-Milgram lemma (see introduction to Section 3.4). □

3.4.50 Theorem (complete existence theorems)**.** *Let Ω be a bounded domain in $\mathbb{R}^d$ with C^1 boundary, and assume the the sesquilinear form $B(u,v)$ is coercive over $H^1(\Omega)$. Let $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$.*

Then the generalized Neumann boundary value problem, find $u \in H^1(\Omega)$ which satisfies

$$B(u,v) \;=\; \int_\Omega v^* f\, dx + \int_{\partial\Omega} v^* g\, ds \tag{3.4.19}$$

for every $v \in C^\infty(\bar{\Omega})$, is uniquely solvable if and only if $u = 0$ is the only solution of the homogeneous equation

$$B(u, v) \;=\; 0$$

for all $v \in C^\infty(\bar{\Omega})$.

More generally, the generalized Neumann boundary value problem, find $u \in H^1(\Omega)$ which satisfies

$$B_\lambda(u, v) \;\equiv\; B(u, v) + \lambda \int_\Omega v^* u \, dx \;=\; \int_\Omega v^* f \, dx + \int_{\partial\Omega} v^* g \, ds \tag{3.4.20}$$

for every $v \in C^\infty(\bar{\Omega})$, is always uniquely solvable except for a countable set of λ's in $\mathbb{C}$. This exceptional set of eigenvalues, λ_j where $j \in \mathbb{N}$, satisfies $\lambda_j \to \infty$ as $j \to \infty$.

When $\lambda = \lambda_j$ in (3.4.20) for some j, the vector space of solutions u to the homogeneous equation,

$$B(u, v) + \lambda_j \int_\Omega v^* u \, dx \;=\; 0 \tag{3.4.21}$$

for every $v \in C^\infty(\bar{\Omega})$, is a finite dimensional subspace of $H^1(\Omega)$.

The proof of this result is similar in spirit to the proof of Theorem 3.4.35 and will be omitted.

3.4.7 Impedance boundary conditions

In this section we consider the problem of finding $u \in H^1(\Omega)$ which satisfies

$$A_\lambda u \;=\; Au + \lambda u \;=\; -\sum_{i,j=1}^{d} \partial_i(a_{ij}\partial_j u) + \sum_{i=1}^{d} b_i \partial_i u + cu + \lambda u \;=\; f \tag{3.4.22}$$

on the domain Ω and

$$Ru \;=\; \frac{\partial u}{\partial \tilde{n}} + \alpha(x) u \;=\; g \tag{3.4.23}$$

on $\partial\Omega$, where the co-normal derivative is

$$\frac{\partial u}{\partial \tilde{n}} \;=\; \sum_{i,j=1}^{d} n_i a_{ij} \partial_j u \tag{3.4.24}$$

defined in a neighborhood of $\partial\Omega$, and $\alpha(x)$ is a function on $\partial\Omega$. Here, n_i are the components of the outward unit normal on $\partial\Omega$ and λ is a real or complex scalar. As usual, all coefficient functions a_{ij}, b_i, c, and α, are in $L^\infty(\Omega)$.

The approach to solving this boundary problem is similar to the preceding section except that now we must include with our Dirichlet form, B, another sesquilinear form over $\partial\Omega$.

3.4.51 Definition. A function $u \in H^1(\Omega)$ will be called a *generalized solution of the impedance boundary value problem* (3.4.22)-(3.4.23) if it satisfies

$$\int_\Omega \partial_i v^* a_{ij} \partial_j u + v^* b_i \partial_i u + v^* c u + \lambda v^* u \, dx + \int_{\partial\Omega} v^* \alpha u \, ds \;=\; \int_\Omega v^* f \, dx + \int_{\partial\Omega} v^* g \, ds \tag{3.4.25}$$

for all $v \in C^\infty(\overline{\Omega})$.

For short we denote the first integral in (3.4.25) by $B_\lambda(u, v)$. Before stating theorems about the existence of solutions to (3.4.25) we will motivate the preceding definition with the following result.

3.4.52 Lemma. *Let $a_{ij} \in C^1(\bar{\Omega})$, $b_i, c, \alpha \in C(\bar{\Omega})$, and $\partial\Omega \in C^1$. If $u \in C^2(\bar{\Omega})$, $f \in C(\bar{\Omega})$, and $g \in C(\partial\Omega)$, then u satisfies (3.4.22) and (3.4.23) if and only if u satisfies (3.4.25) for all $v \in C^\infty(\bar{\Omega})$.*

Proof. $\Rightarrow$) Since the coefficients of A and the boundary differential operator are smooth enough, equations (3.4.22) and (3.4.23) are equations in the vector spaces $C(\bar{\Omega})$ and $C(\partial\Omega)$, respectively. In order that the coefficients of the co-normal derivative be continuous this requires that the normal components n_i be continuous; this is where we use the assumption that $\partial\Omega$ has continuously turning tangent vectors. Therefore (3.4.22) implies that $\int_\Omega v^* A_\lambda u \, dx = \int_\Omega v^* f \, dx$ for every $v \in C^\infty(\bar{\Omega})$. All expressions are smooth enough to invoke Lemma (3.4.45), and we conclude that (3.4.22) implies

$$\int_\Omega \sum_{i,j=1}^{d} \partial_i v^* a_{ij} \partial_j u + v^* \sum_{i=1}^{d} b_i \partial_i u + v^* c u + \lambda u \, dx - \int_{\partial\Omega} v^* \frac{\partial u}{\partial \tilde{n}} \, ds \;=\; \int_\Omega v^* f \, dx \tag{3.4.26}$$

for every $v \in C^\infty(\bar{\Omega})$.

Now (3.4.23) implies that

$$\int_{\partial\Omega} v^* \frac{\partial u}{\partial \tilde{n}} \, ds + \int_{\partial\Omega} v^* \alpha u \, ds \;=\; \int_{\partial\Omega} v^* g \, ds \tag{3.4.27}$$

for every $v \in C^\infty(\bar{\Omega})$, where of course *only* the boundary values of v play a role. Equation (3.4.25) now follows by adding (3.4.26) and (3.4.27).

$\Leftarrow$) As in the first part of the proof we know to begin with that each of the functions f and $A_\lambda u$ are continuous on $\bar{\Omega}$, and g and $\partial u/\partial\tilde{n} + \alpha u$ are continuous on $\partial\Omega$. Lemma (3.4.45) again holds, so (3.4.25) can be written as

$$\int_\Omega v^* A_\lambda u \, dx + \int_{\partial\Omega} v^* \frac{\partial u}{\partial \tilde{n}} + v^* \alpha u \, ds \;=\; \int_\Omega v^* f \, dx + \int_{\partial\Omega} v^* g \, ds. \tag{3.4.28}$$

Now we argue as in the fundamental lemma in the calculus of variations. To show that (3.4.22) holds, we assume that $f \neq A_\lambda u$ at some point of Ω. Since f and $A - \lambda u$ are both continuous this means that there is some open ball in Ω in which the real or imaginary part of $f - A_\lambda u$ is not zero. We can choose for v a smooth function which is zero everywhere except in the neighborhood where $f - Au - \lambda u \neq 0$, and in that neighborhood we choose v to be real valued and have the same sign as the real or imaginary part of $f - Au - \lambda u$. Then the real or imaginary part of $\int_\Omega v^*(f - A - \lambda u) \, dx$ a strictly positive number. This contradicts our assumption and so implies (3.4.22).

The proof of (3.4.23) proceeds in the same way. From what we have just shown the integrals over Ω in (3.4.28) are equal for all $v \in C^\infty(\bar{\Omega})$, and so may be cancelled from both sides to leave $\int_{\partial\Omega} v^* \frac{\partial u}{\partial \tilde{n}} + v^* \alpha u \, ds = \int_{\partial\Omega} v^* g \, ds$ for all $v \in C^\infty(\bar{\Omega})$. If $g - \frac{\partial u}{\partial \tilde{n}} + \alpha u$ had a positive or negative, real or imaginary part in some subset $S \subset \partial\Omega$, we could multiply by a v with corresponding positive or negative, real or imaginary part in that component, and concentrated only in S. This would imply that $\int_{\partial\Omega} v^*(g - \frac{\partial u}{\partial \tilde{n}} + \alpha u) \, ds \neq 0$ for this particular v. Thus, (3.4.23) must hold. □

We now state the solvability theorems.

3.4.53 Theorem (coercive existence theorem)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with sufficiently smooth boundary. Let $\alpha(x)$ be non-negative definite at all points $x \in \partial\Omega$, and let $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$. Then there is a $\lambda_0 \geq 0$ such that the generalized impedance boundary value problem, find $u \in H^1(\Omega)$ which satisfies*

$$\tilde{B}_\lambda(u, v) \;\equiv\; B(u, v) + \lambda \int_\Omega v^* u \, dx + \int_{\partial\Omega} v^* \alpha(x) u \, ds \;=\; \int_\Omega v^* f \, dx + \int_{\partial\Omega} v^* g \, ds \tag{3.4.29}$$

for every $v \in H^1(\Omega)$, has a unique solution whenever $\lambda \geq \lambda_0$.

Proof. The proof is similar to the proof of Theorem 3.4.49. Since B is coercive over H^1 there is a $\lambda_0 \geq 0$ such that B_λ is strictly coercive over H^1 for all $\lambda \geq \lambda_0$.

Since α is non-negative definite for every $x \in \partial\Omega$ the complex number $u(x)^* \alpha(x) \, u(x)$ is non-negative for all $x \in \partial\Omega$ and all $u \in H^1$. Therefore $\int_{\partial\Omega} u^* \alpha(x) u \, ds \;\geq 0$, and $\tilde{B}_\lambda$ is strictly coercive over H^1 when $\lambda \geq \lambda_0$.

We now apply the Lax-Milgram lemma (introduction in Section 3.4) and Lemma 3.4.48. □

The next result relaxes the assumptions that λ be sufficiently large, and that α be non-negative definite.

3.4.54 Theorem (complete existence theorems)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with sufficiently smooth boundary, $f \in L^2(\Omega)$, and $g \in L^2(\partial\Omega)$.*

Then the generalized impedance boundary value problem, find $u \in H^1(\Omega)$ which satisfies

$$\tilde{B}_\lambda(u,v) \;\equiv\; B(u,v) + \lambda \int_\Omega v^* u \, dx + \int_{\partial\Omega} v^* \alpha(x) u \, ds \;=\; \int_\Omega v^* f \, dx + \int_{\partial\Omega} v^* g \, ds \tag{3.4.30}$$

for every $v \in C^\infty(\bar{\Omega})$, is uniquely solvable except for a countable set of λ's in $\mathbb{C}$. This exceptional set of eigenvalues, λ_j where $j \in \mathbb{N}$, satisfies $\lambda_j \to \infty$ as $j \to \infty$.

When $\lambda = \lambda_j$ in (3.4.30) for some j, the vector space of solutions u to the homogeneous equation,

$$\tilde{B}_\lambda(u,v) \;=\; 0 \tag{3.4.31}$$

for every $v \in H^1(\Omega)$, is a finite dimensional subspace of $H^1(\Omega)$.

The proof of this result is similar in spirit to the proof of Theorem 3.4.35 and will be omitted.

3.4.8 Mixed boundary conditions

Frequently in applications one must apply Dirichlet boundary conditions on one portion of $\partial\Omega$ and Neumann or other boundary conditions to the remainder of $\partial\Omega$. When a system of equations is involved one may be applying Dirichlet boundary conditions so some components of u and Neumann, impedance, or oblique derivative conditions to the other components. These problems can be solved by using vector spaces of solutions which are between $H_0^1(\Omega)$ and $H^1(\Omega)$.

We let $\partial\Omega = S_1 \cup S_2$ be the union of two disjoint surfaces S_1 and S_2. We want to solve the problem: find u (in H^1 at least) which solves

$$Au = f \;\text{ on } \Omega, \qquad u = g_1 \;\text{ on } S_1 \qquad Ru = g_2 \;\text{ on } S_2$$

where f, g_1, and g_2 are given functions on Ω, S_1, and S_2, respectively, and where R is the impedance boundary operator (3.4.23). We begin with the case $g_1 = 0$ on S_1.

3.4.55 Definition ($C^\infty_{(0)}$)**.** And let S_1 be a subset of $\partial\Omega$ with non-empty interior relative to $\partial\Omega$. Denote by $C^\infty_{(0)}(\bar{\Omega})$ the subset of $C^\infty(\bar{\Omega})$ consisting of those smooth functions v which satisfy $v = 0$ on S_1. We define the Hilbert subspace $V \subset H^1(\Omega)$ to be the completion of $C^\infty_{(0)}(\bar{\Omega})$ with respect to the norm $\|\cdot\|_1$ of $H^1(\Omega)$.

Since

$$C_0^\infty(\Omega) \subset C^\infty_{(0)}(\bar{\Omega}) \subset C^\infty(\bar{\Omega})$$

we have

$$H_0^1(\Omega) \subset V \subset H^1(\Omega),$$

each Hilbert space in the second line being the completion, with respect to the same norm, of the vector space of smooth functions above it.

3.4.56 Lemma. *Let $\Omega \subset \mathbb{R}^d$ be a bounded open set with $\partial\Omega \in C^1$; and let S_1 be a subset of $\partial\Omega$ with non-empty interior relative to $\partial\Omega$ and relative compliment $S_2 = \partial\Omega \backslash S_1$. Let A be given by (3.4.1) and have coefficient functions $a_{ij}(x) \in C^1(\bar{\Omega})$, and $b_i(x), c(x) \in C(\bar{\Omega})$. Let the Dirichlet sesquilinear form B be given by (3.4.3) and the impedance operator R be given by (3.4.23). Finally, let $u \in C^\infty_{(0)}(\bar{\Omega})$, $f \in C(\bar{\Omega})$, and $g \in C(\overline{S_2})$. Then u satisfies*

$$Au \;=\; f \;\; \text{on } \Omega, \qquad Ru \;=\; g \;\; \text{on } S_2, \qquad \text{and} \qquad u \;=\; 0 \;\; \text{on } S_1 \tag{3.4.32}$$

if and only if

$$B(u,v) + \int_{S_2} v^* \big(Ru - \frac{\partial u}{\partial \tilde{n}}\big)\, ds \;=\; \int_\Omega v^* f \, dx + \int_{S_2} v^* g \, ds \tag{3.4.33}$$

holds for all $v \in C^\infty_{(0)}(\bar{\Omega})$.

Proof. To prove that (3.4.32) implies (3.4.33) we observe that, under the assumptions of the Proposition, the equations in (3.4.32) are equalities in the vector space of continuous functions, either on Ω, S_1, or S_2. So (3.4.32) implies that

$$\int_\Omega v^* Au\,dx = \int_\Omega v^* f\,dx \qquad \text{and} \qquad \int_{S_2} v^* Ru\,ds = \int_{S_2} v^* g\,ds \tag{3.4.34}$$

for all $v \in C_1^\infty$. Under our smoothness assumptions, Lemma 3.4.45 is valid, so

$$\int_\Omega v^* A(x)u\,dx = B(u,v) - \int_{\partial\Omega} v^* \frac{\partial u}{\partial \tilde{n}}\,ds = B(u,v) - \int_{S_2} v^* \frac{\partial u}{\partial \tilde{n}}\,ds \tag{3.4.35}$$

for all $v \in C_1^\infty$, since $v = 0$ on S_1. If we add both equations in (3.4.34), and substitute (3.4.35) for $\int_\Omega v^* Au\,dx$, we obtain (3.4.33).

To prove that (3.4.33) implies (3.4.32) we first observe that Lemma 3.4.45 is again valid, so (3.4.35) is still valid. So (3.4.33) implies that

$$\int_\Omega v^* Au\,dx + \int_{S_2} v^* Ru\,ds \;=\; \int_\Omega v^* f\,dx + \int_{S_2} v^* g\,ds \tag{3.4.36}$$

holds for all $v \in C_1^\infty$. Since $C_0^\infty \subset C_1^\infty$, (3.4.36) implies that $Au = f$ on Ω. So the two complex numbers $\int v^* Au\,dx$ and $\int v^* f\,dx$ are always equal, for any $v \in C_1^\infty$, and may be cancelled in (3.4.36) to give

$$\int_{S_2} v^* Ru\,ds \;=\; \int_{S_2} v^* g\,ds \tag{3.4.37}$$

for all $v \in C_1^\infty$.

Now to show that the two continuous functions Ru and g are equal on the set S_2 it suffices to show they are equal on the interior, S_2°, of S_2 relative to $\partial\Omega$. For the boundary of S_2 in $\partial\Omega$ has measure zero.[20] If they are not equal they must differ at least in the real or imaginary part of at least one component, say the j-th, at some $x_0 \in S_2^\circ$. By continuity, the real part of the j-th component, $\Re\,(Ru-g)_j \geq \epsilon > 0$, for instance, on some δ-neighborhood, $B(x_0,\delta) \cap S_2^\circ$ in S_2°. Here we use the notation $B(x_0,\delta) = \{x \in \mathbb{R}^d\,;\,|x-x_0| < \delta\}$ for the open ball in $\mathbb{R}^d$.

Now, to obtain the same kind of contradiction as in the calculus of variations, take a function $\phi \in C_0^\infty(\mathbb{R}^d;\mathbb{R})$ with support in $B(x_0,\delta)$, and satisfying $0 \leq \phi(x) \leq 1$ for all $x \in \mathbb{R}^d$ and $\phi \equiv 1$ on $B(x_0,\delta/2)$. And construct the function $v \in C^\infty(\bar{\Omega})$ by letting v be zero in all components except the j-th; and in that component set $v_j = \phi$. Then $v \in C^\infty(\bar{\Omega})$. And by its construction we see that

$$\Re \int_{S_2} v^*\big(Ru-g\big)\,ds \;\geq\; \epsilon\,\mathrm{vol}(B(x_0,\delta/2)) \;>\; 0$$

contradicting (3.4.37). □

3.4.57 Definition. Let V be the Hilbert space of Definition 3.4.55. A function $u \in V$ will be called a *generalized solution* of the boundary value problem (3.4.32) if it satisfies (3.4.33) for all $v \in V$.

3.4.58 Theorem. *Let $\Omega \subset \mathbb{R}^d$ be bounded with $\partial\Omega$ smooth. Let A be elliptic, $a_{ij} \in C^1(\overline{\Omega})$, and b_i, and c belonging to $L^\infty(\Omega)$. And let R be given by (3.4.23) and assume that the functions β_i's and α belong to $L^\infty(\partial\Omega)$. Finally, assume that the sesquilinear form*

$$(u,v) \mapsto \int_{\partial\Omega} v^*(R - \frac{\partial}{\partial\tilde{n}})u\,ds \tag{3.4.38}$$

is coercive over V. Then the generalized boundary value problem (3.4.33) is uniquely solvable for $u \in V$ given any $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ except for a discrete set of eigenvalues $\lambda_j \in \mathbb{C}$, $j \in \mathbb{N}$.

The proof of this result is similar in spirit to the proof of Theorem 3.4.35 and will be omitted.

[20] The proof now proceeds just as the end of the proof of Proposition 3.4.46.

3.5 Parabolic and Hyperbolic Equations in the Self-Adjoint Case

Self-adjoint elliptic operators are probably the most common elliptic operators in applications, and the theory for the corresponding parabolic and hyperbolic equations is relatively simple and gives explicit formulas. For simplicity we will only consider homogeneous Dirichlet boundary conditions.

In this section we will be working with functions of t and x. For some $T > 0$ the time interval of interest will be $0 < t < T$. As in earlier sections, $x \in \Omega$, an open subset of $\mathbb{R}^d$. We need the following function spaces.

3.5.1 Definition. We denote by $C^\infty_{(0)}(\Omega \times [0,T])$ the vector space of $C^\infty(\overline{\Omega \times (0,T)})$ functions $f(x,t)$ such that $f(\cdot,t) \in C^\infty_0(\Omega)$ for each $t \in [0,T]$. Define $L^2(0,T;H^1_0(\Omega))$ as the completion of $C^\infty_{(0)}(\Omega \times [0,T])$ with respect to the (Hilbert space) norm

$$\int_0^T \|f(\cdot,t)\|_1^2 \, dt \;=\; \int_0^T \int_\Omega \; |f|^2 + \sum_1^d |\partial_i f|^2 \; dx\, dt\,.$$

$L^2(0,T;H^1_0(\Omega))$ may also be described as the vector space of (measurable) functions $f(x,t)$ on $\Omega \times (0,T)$ such that $f(x,t) = 0$ when $x \in \partial\Omega$, $\partial_j f$ exists as a (measurable) function for each $j = 1, ..., d$, and $\int_\Omega \int_0^T |f(x,t)|^2 + \sum |\partial_j f(x,t)|^2 \, dx\, dt < \infty$. It is a Hilbert space with the inner product

$$(f,g) \;\mapsto\; \int_\Omega \int_0^T f\bar{g} + \sum_1^d \partial_i f \, \partial_i \bar{g} \; dx\, dt\,.$$

3.5.2 Definition. $L^2(0,T;L^2(\Omega))$ is the vector space $L^2(\Omega \times (0,T))$. Its inner product is

$$(f,g) \;\mapsto\; \int_\Omega \int_0^T f\bar{g}\, dx\, dt\,.$$

$L^2(0,T;H^{-1}(\Omega))$ is, by definition, the dual vector space of $L^2(0,T;H^1_0(\Omega))$. To be more explicit, it is the vector space of (finite) linear combinations of $L^2(\Omega \times (0,T))$ functions and 'formal' spatial partial derivatives, $\partial_j g$, such that $g \in L^2(\Omega \times (0,T))$.

Denoting (as usual) by $\|\cdot\|_1$, $\|\cdot\|_0$, and $\|\cdot\|_{-1}$ the norms of $H^1_0(\Omega)$, $L^2(\Omega)$, and $H^{-1}(\Omega)$, respectively, the norms of these three spaces may also be written as the square roots of

$$\int_0^T \|f(\cdot,t)\|_1^2 \, dt\,, \qquad \int_0^T \|f(\cdot,t)\|_0^2 \, dt\,, \qquad \text{and} \qquad \int_0^T \|f(\cdot,t)\|_{-1}^2 \, dt\,.$$

If any of these Hilbert spaces over Ω have an orthonormal basis w_j, these (squared) norms are also equal to

$$\sum_{j=1}^\infty \int_0^T |\alpha_j(t)|^2 \, dt$$

where $f(\cdot,t) = \sum_{j=1}^\infty \alpha_j(t) w_j$ is convergent in the Hilbert space for each $t \in (0,T)$ (Parseval's theorem).

3.5.3 Proposition. *Let Ω and A satisfy the assumptions of Theorem 3.4.35: Let $\Omega \subset \mathbb{R}^d$ be a bounded open subset and A be given by (3.4.1) or (3.4.20). Assume that the coefficient functions $a_{ij}(x)$, $b_i(x)$, and $c(x)$ of the differential operator A belong to $L^\infty(\Omega)$, and that A is elliptic, i.e., satisfies (3.4.5).*

And assume (by integrating by parts) that

$$\int_\Omega \overline{v}\, Au \, dx \;=\; \int_\Omega \overline{Av}\, u \, dx \tag{3.5.1}$$

holds for all $u, v \in C^\infty_0(\Omega)$. Then the eigen-values λ_j and eigen-functions w_j of A, $j \in \mathbb{N}$, established in Theorem 3.4.35, satisfy

a) $\lambda_j \in \mathbb{R}$ *for all* $j \in \mathbb{N}$,

b) $\lambda_j \to +\infty$ *as* $j \to \infty$,

c) $w_j \in H_0^1(\Omega)$ *for all* $j \in \mathbb{N}$, *and*

d) $\int_\Omega \bar{w}_i \, w_j \, dx = 0$ *whenever* $\lambda_i \neq \lambda_j$.

Moreover, the set of eigen-functions can be made to form a complete orthonormal set (hence a basis) for $L^2(\Omega)$.

The reader may check that A will satisfy (3.5.1) if $b_i(x) = 0$ for $i = 1, ..., d$, $c(x)$ is real, and for each x the matrix $[a_{ij}(x)]$ is Hermitian.

Proof. Recall our notation: $\langle u, v\rangle_0 = \int_\Omega \bar{v}\, u \, dx$ and $\|u\|_0^2 = \int_\Omega |u|^2 \, dx$. Assume A satisfies (3.5.1) and that $Aw = \lambda w$. We may replace w by $w/\|w\|_0$ and also assume that $\|w\|_0 = 1$.

Must show that (3.5.1) (has meaning and) 'holds' (by continuity) for all $u, v \in H_0^1(\Omega)$. The meaning is

$$\langle Au, \bar{v}\rangle \;=\; \langle u, \overline{Av}\rangle \tag{3.5.2}$$

where brackets now denote duality pairing.

Each $w_j \in H_0^1(\Omega)$ because the equation $Aw = \lambda_j w_j \in H^{-1}(\Omega)$ has a unique solution $w \in H_0^1(\Omega)$ which is necessarily $w = w_j$.

Then we have

$$\bar{\lambda} = \bar{\lambda}\langle w, \bar{w}\rangle = \langle w, \overline{\lambda w}\rangle = \langle w, \overline{Aw}\rangle = \langle Aw, \bar{w}\rangle = \lambda\langle w, \bar{w}\rangle = \lambda \, .$$

So λ is real.

We know (b) from Theorem 3.4.35.

Now assume $\lambda_i \neq \lambda_j$. Since we know λ_j is real and A satisfies (3.5.2) we have

$$(\lambda_i - \lambda_j)\,\langle w_i, \bar{w}_j\rangle = \langle \lambda_i w_i, \bar{w}_j\rangle - \langle w_i, \overline{\lambda_j w_j}\rangle = \langle Aw_i, \bar{w}_j\rangle - \langle w_i, \overline{Aw_j}\rangle = 0 \, .$$

Since $\lambda_i - \lambda_j \neq 0$ we must have $\langle w_i, w_j\rangle_0 = \langle w_i, \bar{w}_j\rangle = 0$.

Finally we must show $\{w_j\}$ is complete. In fact I will just indicate the idea and leave technical details to a later draft. (But see Dettman section 2.8.) Let $X \subset L^2(\Omega)$ be the closed linear span of $\{w_j\}$. If $X = L^2(\Omega)$ we're done. If not $X^\perp$ is a closed subspace of $L^2(\Omega)$ and A is a self-adjoint operator on it. Just as we considered the Rayleigh quotient

$$\langle Au, u\rangle / \|u\|_0^2$$

to obtain the first eigen-value, eigen-function pair λ_1, w_1, so we can maximize this quotient to obtain another eigen-value and eigen-function in $X^\perp$. Since we have assumed this process has already been carried out and completed in Theorem 3.4.35, we must have $X^\perp = 0$ and $X = L^2(\Omega)$. □

In this section we assume that A satisfies the conditions of Proposition 3.5.3.

3.5.4 Lemma. *Under the preceding assumptions on* A, *a function or distribution* $f = \sum_1^\infty \gamma_j w_j$ *belongs to*

a) $H_0^1(\Omega)$ *if and only if* $\sum_1^\infty \lambda_j |\gamma_j|^2 < \infty$,

b) $L^2(\Omega)$ *if and only if* $\sum_1^\infty |\gamma_j|^2 < \infty$, *and*

c) $H^{-1}(\Omega)$ *if and only if* $\sum_1^\infty \lambda_j^{-1} |\gamma_j|^2 < \infty$.

Conversely, if $f \in H^{-1}(\Omega)$, *then* $f = \sum_1^\infty \gamma_j w_j$ *where* $\gamma_j = \overline{f(w_j)} = \overline{\langle f, w_j\rangle}$.

Proof. We use the fact that $C_0^\infty(\Omega) \hookrightarrow H_0^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$ with every embedding dense. The theory of compact self-adjoint operators on $L^2(\Omega)$ applies to A^{-1}; it tells us that the eigen-functions w_j of A form a complete orthonormal set (a basis) for $L^2(\Omega)$, and that the eigen-values λ_j are real, positive, and $\to +\infty$ as $j \to \infty$. Since A is a bijection no λ_j is zero. Also since A is a bijection and $\lambda_j w_j \in H^{-1}(\Omega)$, the unique solution w_j of $Au = \lambda_j w_j \in H^{-1}(\Omega)$ must lie in $H_0^1(\Omega)$.

Now, we already know from Parseval's equality that an $L^2(\Omega)$ function $f = \sum_1^\infty \gamma_j w_j$, convergent in $L^2(\Omega)$, satisfies $\|f\|^2 = \sum_1^\infty |\gamma_j|^2$. And conversely, if $\sum_1^\infty |\gamma_j|^2 < \infty$ then $f_n = \sum_1^n \gamma_j w_j$ is Cauchy in $L^2(\Omega)$, hence $f = \lim f_n$ exists in $L^2(\Omega)$.

THIS PROOF IS TO BE FINISHED (Treves BASIC). □

The Static Boundary Value Problem We pause here to write out the solution of the elliptic boundary value problem when A is a strictly positive, self-adjoint operator on $L^2(\Omega)$.[21] We know that $A : H_0^1(\Omega) \to H^{-1}(\Omega)$ is a bijection and that any solution u of the equation $Au = f$, which lies in $H_0^1(\Omega)$, will automatically satisfy homogeneous Dirichlet boundary conditions.

3.5.5 Proposition. *Let $f = \sum_1^\infty \gamma_j w_j \in H^{-1}(\Omega)$ where $\gamma_j = \overline{\langle f, w_j \rangle}$ as in Lemma 3.5.4. Then the equation $Au = f$ is solved by $u = \sum_1^\infty \nu_j\, w_j$ where $\nu_j = \gamma_j/\lambda_j$ and the series converges in $H_0^1(\Omega)$.*

Proof. Since the w_j's are orthogonal, the equation $\sum_1^\infty \nu_j \lambda_j w_j = \sum_1^\infty \gamma_j w_j$ implies $\nu_j = \gamma_j/\lambda_j$ for each j.

Let's check that the condition $\sum_1^\infty \lambda_j^{-1} |\gamma_j|^2 < \infty$ implies $\sum_1^\infty \lambda_j |\nu_j|^2 < \infty$. In fact this is clear because

$$\sum_1^\infty \lambda_j |\nu_j|^2 \;=\; \sum_1^\infty \lambda_j (|\gamma_j|^2/\lambda_j^2) \;=\; \sum_1^\infty \lambda_j^{-1} |\gamma_j|^2 \,.$$

This tells us that $u \in H_0^1(\Omega)$.

Finally we check that this u satisfies $Au = f$ by verifying that $Au = \sum_1^\infty \nu_j A w_j$. But this holds because $A : H_0^1(\Omega) \to H^{-1}(\Omega)$ is continuous by Lemma 3.4.1. □

Now let's consider some time evolution equations associated with A. NOTE[22]

Parabolic Equations The parabolic equation

$$\partial_t u + Au \;=\; f(x,t)\,, \qquad x \in \Omega,\ 0 < t < T\,, \tag{3.5.3}$$

with initial condition $u(x,0) = u_0(x) \in L^2(\Omega)$ and $f \in L^2(0,T; H^{-1}(\Omega))$, has a unique solution $u \in L^2(0,T; H_0^1(\Omega))$ under our assumptions on A. We compute it by the 'separation of variables' method.

Let

$$f(x,t) \;=\; \sum_1^\infty \gamma_j(t) w_j(x) \qquad \text{and} \qquad u_0(x) \;=\; \sum_1^\infty \alpha_j w_j(x)\,,$$

and assume u has the form $u = \sum_1^\infty \nu_j(t) w_j$ with unknown functions $\nu_j(t)$. Formally applying the differential operator to this series term-by-term we obtain

$$\partial_t u + Au \;=\; \sum_1^\infty (\nu_j' + \lambda_j \nu_j) w_j \;=\; \sum_1^\infty \gamma_j(t) w_j\,.$$

Since the w_j's are linearly independent we have

$$\nu_j'(t) + \lambda_j \nu_j(t) = \gamma_j(t)\,, \qquad j = 1, 2, ...$$

The initial condition requires $\nu_j(0) = \alpha_j$ for each j. These ordinary differential equations have solutions

$$\nu_j(t) \;=\; \alpha_j\, e^{-\lambda_j t} + \int_0^t e^{-\lambda_j (t-s)} \gamma_j(s)\, ds\,. \tag{3.5.4}$$

We can prove

3.5.6 Theorem. *Let $u_0(x) \in L^2(\Omega)$ and $f \in L^2(0,T; L^2(\Omega))$. Then $u = \sum_1^\infty \nu_j(t)\, w_j$, where $\nu_j(t)$ is given by (3.5.4), belongs to $L^2(0,T; H_0^1(\Omega))$, and satisfies (3.5.3) and $\lim_{t\to 0+} u(t) = u_0$ in $L^2(\Omega)$.*

The conclusion of this theorem remains true if it is only assumed that $f \in L^2(0,T; H^{-1}(\Omega))$, however the proof of this stronger statement is rather technical.

[21] Actually, A is not required to be strictly positive (coercive). It is sufficient that no λ_j be zero. Even then we could state the 'alternative.'

[22] I must define the spaces $L^2(0,T; H^k(\Omega))$.

Proof. Assume $\sum_1^\infty \int_0^T |\gamma_j(t)|^2\,dt < \infty$ and $\sum_1^\infty |\alpha_j|^2 < \infty$. We must show that $\sum_1^\infty \lambda_j \int_0^T |\nu_j|^2\,dt < \infty$.

Write $u(x,t) = v(x,t) + w(x,t)$ where $v = \sum_1^\infty \alpha_j e^{-\lambda_j t} w_j$ is the solution for the non-homogeneous initial conditions, and $w = \sum_1^\infty \int_0^t e^{-\lambda_j(t-s)}\gamma_j(s)\,ds\,w_j$ is the solution for the non-homogeneous forcing function. It then suffices to check that both v and w are in $L^2(0,T;H_0^1(\Omega))$, i.e., that

$$\sum_1^\infty \lambda_j \int_0^T |\alpha_j e^{-\lambda_j t}|^2\,dt \;<\; \infty \qquad \text{and} \qquad \sum_1^\infty \lambda_j \int_0^T \Big|\int_0^t e^{-\lambda_j(t-s)}\gamma_j(s)\,ds\Big|^2\,dt \;<\; \infty\,. \tag{3.5.5}$$

Since $\lambda_j > 0$,

$$\int_0^T |e^{-\lambda_j t}|^2\,dt \;=\; \int_0^T e^{-2\lambda_j t}\,dt \;=\; \frac{1-e^{-2\lambda_j T}}{2\lambda_j} \;\le\; \frac{1}{2\lambda_j}\,.$$

So the left side of (3.5.5) is bounded by $\sum_1^\infty \lambda_j\,(|\alpha_j|^2\,\frac{1}{2\lambda_j}) = \frac{1}{2}\|u_0\|_0^2$ which we assume is finite.

Next we use Schwarz' inequality to obtain

$$\begin{aligned}\Big|\int_0^t e^{-\lambda_j(t-s)}\gamma_j(s)\,ds\Big|^2 \;&\le\; \int_0^t e^{-2\lambda_j(t-s)}\,ds\;\int_0^t |\gamma_j(s)|^2\,ds \\ &=\; e^{-2\lambda_j t}\Big[\frac{e^{2\lambda_j s}}{2\lambda_j}\Big]_{s=0}^t \int_0^t |\gamma_j(s)|^2\,ds \;\le\; \frac{1}{2\lambda_j}\int_0^T |\gamma_j(s)|^2\,ds\,.\end{aligned}$$

Since T is finite, the right side of (3.5.5) is now easily obtained.

Finally, we leave it to the reader to show that

$$\|u(\cdot,t) - u_0\|_0^2 \;=\; \sum_1^\infty \Big|\alpha_j e^{-\lambda_j t} + \int_0^t e^{-\lambda_j(t-s)}\,\gamma_j(s)\,ds - \alpha_j\Big|^2 \;\to\; 0$$

as $t \to 0+$. □

Hyperbolic Equations The hyperbolic equation

$$\partial_t^2 u + Au \;=\; f(x,t)\,, \qquad x \in \Omega,\ 0 < t < T\,, \tag{3.5.6}$$

with initial conditions

$$u(x,0) = u_0(x) \in H_0^1(\Omega) \qquad \text{and} \qquad \partial_t u(x,0) = u_1(x) \in L^2(\Omega)\,, \tag{3.5.7}$$

and forcing function

$$f \in L^2(0,T;L^2(\Omega))\,, \tag{3.5.8}$$

has a unique solution $u \in L^2(0,T;H_0^1(\Omega))$ under the same assumptions on A as in the parabolic case. Again, it can be computed by using the eigen-functions of A and separating the time and space variables.

Let

$$f(x,t) \;=\; \sum_1^\infty \gamma_j(t)w_j(x)\,, \qquad u_0(x) \;=\; \sum_1^\infty \alpha_j w_j(x)\,, \qquad \text{and} \quad u_1(x) \;=\; \sum_1^\infty \beta_j w_j(x)\,.$$

Assume u has the form $u = \sum_1^\infty \nu_j(t)w_j$ with unknown functions $\nu_j(t)$. Again proceeding formally as in the parabolic case, we apply the differential operator to this series term-by-term to obtain

$$\partial_t^2 u + Au \;=\; \sum_1^\infty (\nu_j'' + \lambda_j\,\nu_j)w_j \;=\; \sum_1^\infty \gamma_j(t)w_j\,.$$

The linear independence of the w_j's imply that

$$\nu_j''(t) + \lambda_j\,\nu_j(t) \;=\; \gamma_j(t)\,, \qquad j = 1,2,....$$

The initial conditions require that $\nu_j(0) = \alpha_j$ and $\nu_j'(0) = \beta_j$ for each j. These ordinary differential equations have solutions

$$\nu_j(t) \;=\; \alpha_j \cos(\sqrt{\lambda}_j t) \;+\; \beta_j \frac{\sin(\sqrt{\lambda}_j t)}{\sqrt{\lambda}_j} \;+\; \int_0^t \frac{\sin[\sqrt{\lambda}_j (t-s)]}{\sqrt{\lambda}_j}\, \gamma_j(s)\, ds\,. \tag{3.5.9}$$

We can prove the following

3.5.7 Theorem. *Assume that $\Omega \subset \mathbb{R}^d$ is bounded and A is a self-adjoint, strictly positive elliptic operator, as in THEOREM TBD, with eigen-values and eigen-functions given by (??). Then for any u_0, u_1, and f satisfying (3.5.7) and (3.5.8), there is a unique solution $u \in L^2(0,T; H_0^1(\Omega))$ of (3.5.6) and it has the form $u = \sum_1^\infty \nu_j(t) w_j$ with ν_j given by (3.5.9). NOTE*[23]

Proof. We must use (3.5.9) to show that

$$\sum_1^\infty \lambda_j \int_0^T |\nu_j(t)|^2\, dt < \infty \tag{3.5.10}$$

where we get to assume that

$$\sum_1^\infty \lambda |\alpha_j|^2 < \infty\,, \qquad \sum_1^\infty |\beta_j|^2 < \infty\,, \qquad \text{and} \qquad \sum_1^\infty \int_0^T |\gamma_j(t)|^2\, dt < \infty\,. \tag{3.5.11}$$

We first use the inequality $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ valid for any real numbers a, b, c. This is seen by expanding the left side and using $2ab \le a^2 + b^2$, which holds because $0 \le (a-b)^2$. Thus, for each t,

$$\lambda_j |\nu_j(t)|^2 \;\le\; 3\Big\{\lambda_j |\alpha_j|^2 + \lambda_j (|\beta_j|^2/\lambda_j) + \lambda_j \Big[\int_0^t \sin[\sqrt{\lambda}_j (t-s)]\, \gamma_j(s)\, ds\Big]^2 / \lambda_j \Big\}\,.$$

Using the Cauchy-Schwarz's inequality on the integral shows that the last term is bounded by

$$\int_0^t 1\, ds \;\int_0^t |\gamma_j(s)|^2\, ds \;\le T \int_0^T |\gamma_j(s)|^2\, ds\,.$$

There is then a constant $C > 0$ such that

$$\lambda_j |\nu_j(t)|^2 \;\le\; C\Big\{\lambda_j |\alpha_j|^2 + |\beta_j|^2 + \int_0^T |\gamma_j(s)|^2\, ds\Big\}, \quad 0 \le t \le T\,.$$

If we integrate on t and sum over j, (3.5.10) follows immediately from (3.5.11).

Uniqueness will be shown in a later draft, or can be shown by the reader. □

3.5.8 Exercise. State and prove an analog of Proposition 3.5.5 when a finite number of the λ_j's are < 0.

3.5.9 Exercise. State and prove an analog of Proposition 3.5.5 when a finite number of the λ_j's are $= 0$. In particular, for which $f \in H^{-1}(\Omega)$ can the equation $Au = f$ be solved? (This will lead to the Fredholm alternative.)

3.5.10 Exercise. State and prove an extension of Proposition 3.5.5 where the equation $Au = f$ is replaced by the equation $(A - \lambda)u = f$ with $\lambda \in \mathbb{C}$. Consider both cases, $\lambda \in \{\lambda_j\ ;\ j = 1, 2, ...\}$ and $\lambda \notin \{\lambda_j\}$.

3.5.11 Exercise. Let $L > 0$. Find an orthonormal basis of eigen-functions of $-\frac{d^2}{dx^2}$ in $L^2(0, L)$. Choose the eigen-functions to be zero at $x = 0$ and $x = L$.

Use these eigen-functions $w_j(x)$ and eigen-values λ_j to give a formula for the solution u of the initial value problem

$$(\partial_t - \partial_x^2) u = f(x,t), \text{ for } 0 < x < L, t > 0, \quad \text{and} \quad u(x,0) = u_0(x)\,.$$

[23] I have yet to make precise sense of how u is related to the initional conditions, and prove it.

Do the same for the equation

$$(\partial_t^2 - \partial_x^2)u \;=\; f(x,t), \quad \text{for} \quad 0 < x < L, t > 0$$

with initial values

$$u(x,0) \;=\; u_0(x), \quad \text{and} \quad \partial_t u(x,0) \;=\; u_1(x)\,.$$

3.5.12 Exercise. With the same set-up as in the last exercise, give a formula for the solution u of the equation $(\partial_t^2 + b\,\partial_t - \partial_x^2)u = f$ with initial values $u(x,0) = u_0(x)$ and $\partial_t u(x,0) \;=\; u_1(x)$. Assume $b \in \mathbb{R}$ and $\neq 0$. (The term $b\,\partial_t u$ models looses such as electrical resistance, and has been used to model signal propagation over long distances in a wire. Therefore, this equation is called the telegraph equation.)

When $f = 0$ but u_0 and/or u_1 are not zero, for what values of b (if any) does $u(\cdot,t) \to 0$ in either $L^2(\Omega)$ or $H_0^1(\Omega)$ as $t \to \infty$? (Note, for example, that $u \to 0$ in $L^2(\Omega)$ means that $\|u(\cdot,t)\|^2_{L^2(\Omega)} \to 0$.

3.5.13 Exercise. Let Ω be a bounded open set in $\mathbb{R}^d$ and A a positive, self-adjoint elliptic operator on $L^2(\Omega)$ with complete orthonormal set of eigen-functions w_j in $L^2(\Omega)$ and eigen-values $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_j \leq \cdots \to +\infty$. Let $R : L^2(\Omega) \to L^2(\Omega)$ be a bounded linear transformation given by $Rw_j = b_j w_j$ for $j \in \mathbb{N}$, where the sequence $b_j \in [0,\infty)$ is bounded. Assume that $f = f(x,t) = \sum_1^\infty \gamma_j(t)\, w_j(x)$ as before. Give a formula for the solution $u(x,t)$ of the initial value problem

$$(\partial_t^2 + R\,\partial_t + A)u \;=\; f(x,t), \quad \text{for} \quad x \in \Omega, t > 0$$

with

$$u(x,0) \;=\; u_0(x), \quad \text{and} \quad \partial_t u(x,0) \;=\; u_1(x)\,.$$

3.5.14 Exercise. Let Ω, A, and f be as in the last exercise. Give a formula for the solution u of the Schrödinger equation

$$(-i\,\partial_t + A)u = f(x,t), \quad \text{for} \quad x \in \Omega, t > 0, \quad \text{with} \quad u(x,0) = u_0(x)\,.$$

Typically, for a model of n electrons moving in 3-space, the highest order term of A is the Laplacian $-\Delta$ on $\mathbb{R}^d$ with $d = 3n$.

4 Calculus of Variations

The calculus of variations is about finding the function $u \in C^m(\Omega; \mathbb{R}^r)$ that optimizes a functional of the form

$$F(u) \;=\; \int_\Omega f(x, u(x), Du(x), D^2u(x), ..., D^m u(x))\, dx \;. \tag{4.0.1}$$

Here $\Omega \subset \mathbb{R}^d$, f is a real-valued function of $d + r + dr + d^2r + \cdots + d^m r$ variables, and $D^k u$ stands for all k-th order derivatives of u. For example, if $f(x, y, z) = \sqrt{1 + z^2}$ is a function of three (but only one in this case) real variables the functional

$$F(u) \;=\; \int_a^b \sqrt{1 + (u'(x))^2}\, dx \tag{4.0.2}$$

is the arclength of the curve $(x, u(x))$ over the interval $a \le x \le b$.

When optimizing a function of d variables in calculus the first derivative test yields d equations in the d unknowns. The first derivative test for functionals of the form (4.0.1) yields a differential equation which must still be solved for the unknown optimizer u. The existence theory for such equations is not trivial and will not be addressed here. In fact, we will show that it is easy to describe problems of the form (4.0.1) which have no solution, at least in the classical sense.

The first derivative test requires only the notion of Gateaux (directional) derivative (section **??**); the second derivative test and the theorem on optimization with constraints uses the Frechet derivative.

4.1 Examples and Preliminary Observations

We begin with some example problems. The solutions to these and others will be given in later sections.

4.1.1 Example (Arc length and surface area)**.** First we generalize the functional (4.0.2). Let $u \in C^1([a, b]; \mathbb{R}^r)$ describe a parameterized curve in $\mathbb{R}^r$. The length of this curve is given by

$$L(u) \;=\; \int_a^b |u'(x)|\, dx$$

where $|u'(x)| = \sqrt{u_1'(x)^2 + \cdots + u_r'(x)^2}$. A natural question is to find the shortest curve which joins two points $y_0 = u(a)$ and $y_1 = u(b)$ in $\mathbb{R}^r$.

Let Ω be an open subset in $\mathbb{R}^d$ and $u \in C^1(\Omega; \mathbb{R})$. The graph of u defines a d dimensional surface in $\mathbb{R}^{d+1}$. The 'area' of this surface is given by

$$A(u) \;=\; \int_\Omega \sqrt{1 + |\nabla u(x)|^2}\, dx$$

where $\nabla u(x)$ is the gradient of u. A natural question is to find the surface of minimal area which takes specified values $u(x) = g(x)$ when $x \in \partial\Omega$ and g is given on $\partial\Omega$. Such surfaces are formed by soap films when a wire is dipped into a bowl of liquid soap.

4.1.2 Example (Particle dynamics in a potential field)**.** We consider n particles in $\mathbb{R}^3$ that are moving under the influence of a potential field U defined on the state space $\mathbb{R}^{3n}$. The potential may depend only on the positions of the n particles, or on these positions and time. Letting $x \in C^2([a, b]; \mathbb{R}^{3n})$ denote the coordinates of the n particles as a function of time, we write $U(x(t), t)$ for the potential and $T(t)$ for the kinetic energy of the particles. The kinetic energy usually has the form $T(t) = \frac{1}{2}\dot{x}(t)^t M\dot{x}(t)$ where M is a diagonal matrix of masses.

Hamilton's principle of stationary action says that the motion of this system, when $a < t < b$, will follow the trajectory $x(t)$ for which the first variation (section **??**) of the *action functional*

$$A(x) \;=\; \int_a^b T(t) - U(x(t), t)\, dt$$

is zero. In some cases this trajectory will minimize the action functional, the *principle of least action*, and in others it will not. We will soon see that such a trajectory always satisfies the Newtonian equations $M\ddot{x}(t) + \nabla U(x(t), t) = 0$ on $a < t < b$.

4.1.3 Example (Brachistochrone)**.** The brachistochrone problem was first described (1630's) by Galileo Galilei, and later (1690's) used by John Bernoulli to challenge mathematicians at the time for a solution. Let (a, A), $(b, B) \in \mathbb{R}^2$ with $a < b$ and $A > B$. And let the planar curve $(x, y(x))$, $a \le x \le b$ with $y(a) = A$ and $y(b) = B$, be the coordinates of a frictionless wire on which a bead slides. The problem is to find the curve $y(x)$ which gives the shortest time of decent for the bead, when it falls (from rest) from point (a, A) to (b, B) under the force of gravity. (The force of gravity is applied in the direction of the negative y-axis.) The decent time for this bead is given by the integral

$$T \;=\; \int_a^b \sqrt{\frac{1 + (y'(x))^2}{2\,g\,y(x)}}\,dx\;, \tag{4.1.1}$$

and Galileo had already observed that a straight line joining the two points does not give the shortest time.

A variation on the brachistochrone problem is the design of a good child's slide by choosing the curve which minimizes (4.1.1) subject to the constraint the $y'(a) = m$ for some slope $m \le 0$. It is thought that if m is too negative the slide will be too scary for children.

4.1.4 Example (Minimal surface of revolution)**.** Let the locus of points $(x, y(x))$, $a \le x \le b$, describe a curve joining the points (a, A) and (b, B) in $\mathbb{R}^2$. The area of the surface of revolution of this curve is

$$\int_a^b 2\pi\, y(x)\, \sqrt{1 + y'(x)^2}\,dx\;.$$

The problem is to find the curve $y(x)$ that gives the least surface area when (a, A) and (b, B) are specified.

4.1.5 Example (Cubic smoothing spline)**.** The following question is important in some data analysis problems. Suppose we are given n data points (x_1, y_1), (x_2, y_2), ...,(x_n, y_n) in $\mathbb{R}^2$. The idea here is that the x_j's are observed precisely but the y_j's can only be observed with some additive, usually random, error. The question is: What is the real-valued function $u(x)$, $a \le x \le b$, which minimizes

$$\lambda \int_a^b (u''(x))^2\,dx + \sum_{j=1}^{n} (y_j - u(x_j))^2\;?$$

The first term in this expression is a least squares term which forces $u(x_j)$ to lie near the data value y_j. The second term penalizes u if it is too rough (its curvature is too great as measured by u'') on the interval (a, b). The important parameter λ is chosen to control the trade-off between these two conflicting requirements.

4.1.6 Example (Cubic interpolating spline)**.** This example is just like the preceding one, except that we assume the y_j's have no error. So we want to force u to interpolate all the points (x_j, y_j) exactly. In this case we choose u to minimize

$$\int_a^b (u''(x))^2\,dx$$

subject to the constraint that

$$u(x_1) \;=\; y_1,\; u(x_2) \;=\; y_2, \qquad \ldots \qquad ,\; u(x_n) \;=\; y_n\;.$$

4.1.7 Example (Econ/Harvest rate)**.**

4.1.8 Example (Vibrating string)**.** This example is similar to Example 4.1.2. Let $u(x, t)$ be the displacement at position $x \in [a, b]$ and time $t \ge 0$ of a taut elastic string when the string is given some initial displacement or velocity. The kinetic energy of this continuum mechanical system is

$$T(t) \;=\; \int_a^b \tfrac{1}{2}\Big(\frac{\partial u}{\partial t}(x, t)\Big)^2\,\rho(x)\,dx$$

where $\rho(x)$ is the density of the string at location x. This expression is the pointwise kinetic energy, at each point x, integrated over the entire length of the string.

The potential energy of this system is given by the tension on the elastic string times the length of the string.[24] If this potential is normalized to zero when the length of the string is $b-a$ (no deflection of the string's position), it may be written

$$U(t) \;=\; \tau \int_a^b \sqrt{1+(\partial u(x,t)/\partial x)^2}\, dx - \tau\,(b-a)$$

where the constant τ is the tension. This potential can be used in an action integral as in Example 4.1.2, but we may derive a simpler 'small displacement approximation' by approximating $U(t)$ by its lowest order term

$$\int_0^T T(t) - U(t)\, dt \;=\; \int_0^T \int_a^b \tfrac{1}{2}\rho\,(\frac{\partial u}{\partial t})^2 - \tfrac{1}{2}\tau\,(\frac{\partial u}{\partial x})^2\, dx\, dt\ , \tag{4.1.2}$$

which follows from $\sqrt{1+z^2}-1 = \frac{1}{2}z^2 + \cdots$ evaluated at $z = \partial u/\partial x$.

4.1.9 Example (Non-existence of extremals)**.** It is easy to see geometrically that some functionals have no minimum. Consider the problem of finding a curve $(x, u(x))$, $a \le x \le b$, of minimum length (4.0.2) subject to the constraint that $u'(a) = 1$ and $u'(b) = -1$. The infimum of (4.0.2) subject to this constraint is $b-a$ but no curve achieves this length.

Consider again the functional (4.0.2). If we restrict the domain of F to those $u \in C^1([a,b])$ subject to the constraints that $u(a) = \alpha$, $u(b) = \beta$, and $u((a+b)/2) = \gamma$ where $\alpha > \beta > \gamma$ it is easy to see that the minimizing length is given by a function with a corner at $(a+b)/2$. So the minimization problem has no solution in $C^1([a,b])$.

Let g be defined on $[0,1]$ by $g(x) = 1$ if $0 \le x \le 1/2$ and $g(x) = -1$ if $1/2 < x \le 1$. Consider now the problem of choosing $u \in C([0,1])$ which minimizes the distance to g, i.e., which minimizes the functional

$$F(u) \;=\; \int_0^1 |g(x) - u(x)|^2\, dx\ .$$

By choosing u appropriately the value of this functional can be made as close to 0 as desired, but 0 is nover achieved for any continuous function u.[25]

The main objective of this section (chapter) will be to derive the 'Euler-Lagrange' differential equations for various functionals. These equations characterize the extremals of each functional, just as the first derivative test does in the calculus of functions on $\mathbb{R}^n$. The derivation of the Euler-Lagrange equations is based on the following

4.1.10 Theorem (Fundamental Lemma of the Calculus of Variations)**.** *Let $\Omega \subset \mathbb{R}^d$ be open and $f : \Omega \to \mathbb{R}$ be continuous. If*

$$\int_\Omega f(x)\phi(x)\, dx \;=\; 0 \tag{4.1.3}$$

for all $\phi \in C_0^\infty(\Omega;\mathbb{R})$ then $f \equiv 0$ on Ω.

Proof. Suppose $f(x_0) \neq 0$ for some $x_0 \in \Omega$. Let's assume $f(x_0) > 0$, otherwise we could consider $-f$. The $\epsilon\delta$-definition of continuity means that there is a ball $B = B_\delta(x_0) \subset \Omega$ such that $|f(x) - f(x_0)| < \epsilon$ if $x \in B$. In particular, $f(x) > f(x_0) - \epsilon$ when $x \in B$. And if we choose, as we may, $\epsilon \le f(x_0)/2$ we have $f(x) \ge f(x_0)/2$ when $x \in B$.

Now we know from Lemma 1.2.3 that there is a $\phi \in C_0^\infty(B_\delta(x_0))$ which satisfies $0 \le \phi \le 1$ and $\phi \equiv 1$ on $B_{\delta/2}(x_0)$.

With these observations in hand and ϕ chosen as above we have shown that the left side of (4.1.3) satisfies

$$\int_B f(x)\phi(x)\, dx \ge \frac{f(x_0)}{2} \int_{B_{\delta/2}(x_0)} dx > 0$$

which is a contradiction. We conclude that $f \equiv 0$ on Ω. □

[24] TO STEVE: Say this better.

[25] Courant, R. and Robbins, *What Is Mathematics?* (1943), ch VII, gives a nice discussion of the calculus of variations problems, especially of minimal surfaces and non-existence/uniqueness.

4.1.11 Exercise. Prove the following variation of Theorem 4.1.10: *If* $\int_\Omega f(x)\phi(x)\,dx = 0$ *for all* $\phi \in C_0(\Omega;\mathbb{R})$ *then* $f \equiv 0$ *on* Ω. Follow the proof above except do not use Lemma 1.2.3; instead construct a simpler continuous function which is 1 on $B_{\delta/2}(x_0)$ but 0 on the compliment of $B_\delta(x_0)$.

4.1.12 Corollary. *Let* $\Omega \subset \mathbb{R}^d$ *be open and* $f : \Omega \to \mathbb{R}^r$ *be continuous. If*

$$\int_\Omega f(x) \cdot \phi(x)\,dx \;=\; 0 \tag{4.1.4}$$

for all $\phi \in C_0^\infty(\Omega;\mathbb{R}^r)$ *then* $f \equiv 0$ *on* Ω.

Proof. By setting all components of ϕ equal to zero on Ω except one, we may apply the lemma successively to each component of f. □

4.1.13 Corollary. *Let* Ω *be a* d *dimensional 'surface' in* $\mathbb{R}^{d+k}$, *some* $k > 0$, *with surface area element* ds. *And let* $f : \Omega \to \mathbb{R}^r$ *be continuous. If*

$$\int_\Omega f(x) \cdot \phi(x)\,ds(x) \;=\; 0 \tag{4.1.5}$$

for all $\phi \in C_0^\infty(\Omega;\mathbb{R}^r)$ *then* $f \equiv 0$ *on* Ω.

Here $C_0^\infty(\Omega;\mathbb{R}^r)$ is the vector space of smooth functions which are zero outside some relatively compact subset of Ω.

Proof. In sufficiently small subsets of Ω the surface area element looks like $ds(x) = \sigma(x)\,dx = \sigma(x_1,...,x_d)\,dx_1...dx_d$ where $x = (x_1,...,x_d)$ are the local coordinates. The area density σ is a strictly positive function on all of Ω. By choosing δ in the proof of the fundamental lemma so small that the ball $B_\delta(x_0)$ can be covered by a single set of coordinates, the proof here is reduced to the same argument as before. The conclusion is that $f(x)\,\sigma(x) \equiv 0$ on Ω, and hence that $f(x) \equiv 0$ as well. □

We will make frequent use of the following result on differentiating under the integral sign. The reader may construct the proof, or consult Rudin's *Principles of Mathematical Analysis*, Marsden's *Elementary Classical Analysis*, or Apostol's *Mathematical Analysis*.

4.1.14 Proposition. *Let* $\Omega \subset \mathbb{R}^d$ *have compact closure and* $[a,b]$ *be a compact interval, and let* $f : \bar\Omega \times [a,b] \to \mathbb{R}$ *be continuous and* $\partial f(x,t)/\partial t$ *also be continuous on* $\bar\Omega \times [a,b]$. *Then* $\phi(t) = \int_\Omega f(x,t)\,dx$ *is differentiable on* $a \le t \le b$ *and*

$$\phi'(t) \;=\; \int_\Omega \frac{\partial f}{\partial t}(x,t)\,dx\ .$$

4.2 Real Valued Functions of a Real Variable

We first consider one of the simplest problems in the calculus of variations.

4.2.1 Theorem. *Assume* $[a,b] \subset \mathbb{R}$ *is compact,* $f \in C^2([a,b]\times\mathbb{R}^2;\mathbb{R})$, *and set* $X = C^2([a,b];\mathbb{R})$. *Let* $F : X \to \mathbb{R}$ *be defined by*

$$F(u) = \int_a^b f(x,u(x),u'(x))\,dx\ . \tag{4.2.1}$$

If $u \in X$ *is a local extremal of* F *then* u *satisfies the differential equation*

$$\frac{d}{dx} f_{u'} - f_u = 0 \tag{4.2.2}$$

or more explicitly

$$f_{xu'} + f_{uu'}u' + f_{u'u'}u'' - f_u = 0.$$

Furthermore, at the end points of the interval, u *satisfies*

$$\frac{\partial f}{\partial u'}(a,u(a),u'(a)) = 0 \tag{4.2.3}$$

and

$$\frac{\partial f}{\partial u'}(b, u(b), u'(b)) = 0 \ . \tag{4.2.4}$$

Equation (4.2.2) is called the *Euler-Lagrange equation* for this variational problem, and the relations (4.2.3) and (4.2.4) are called *natural boundary conditions*.

Proof. If u is an extremal then the first variation $\delta F(u;h) = 0$ for all $h \in X$ by Theorem 1.8.8. Letting $' = \frac{d}{dx}$, and $\partial f/\partial u$ and $\partial f/\partial u'$ denote the partial derivatives of f with respect to its second and third variables, we calculate the variation

$$\begin{aligned}
\delta F(u;h) &= \frac{d}{dt}\left[\int_a^b f(x, u(x) + t\,h(x), u'(x) + t\,h'(x))\,dx\right]_{t=0} \\
&= \int_a^b \frac{d}{dt}\Big[f(x, u(x) + t\,h(x), u'(x) + t\,h'(x))\Big]_{t=0}\,dx \\
&= \int_a^b \Big[\,\frac{\partial f}{\partial u}(x, u(x) + t\,h(x), u'(x) + t\,h'(x))\,h(x) \\
&\qquad\qquad + \frac{\partial f}{\partial u'}(x, u(x) + t\,h(x), u'(x) + t\,h'(x))\,h'(x)\Big]_{t=0}\,dx \\
&= \int_a^b \Big[\,\frac{\partial f}{\partial u}(x, u(x), u'(x))\,h(x) + \frac{\partial f}{\partial u'}(x, u(x), u'(x))\,h'(x)\Big]\,dx \\
&= \int_a^b \frac{\partial f}{\partial u}(x, u(x), u'(x))\,h(x)\,dx + \Big[\,\frac{\partial f}{\partial u'}(x, u(x), u'(x))\,h(x)\Big]_{x=a}^{x=b} \\
&\qquad\qquad - \int_a^b \Big(\frac{d}{dx}\frac{\partial f}{\partial u'}(x, u(x), u'(x))\Big)\,h(x)\,dx
\end{aligned}$$

$$\begin{aligned}
&= \int_a^b \Big[\,\frac{\partial f}{\partial u}(x, u(x), u'(x)) - \frac{d}{dx}\frac{\partial f}{\partial u'}(x, u(x), u'(x))\Big]\,h(x)\,dx \\
&\qquad\qquad + \frac{\partial f}{\partial u'}(b, u(b), u'(b))\,h(b) - \frac{\partial f}{\partial u'}(a, u(a), u'(a))\,h(a).
\end{aligned}$$

The differentiation can be moved inside the integral in the second equality by Proposition 4.1.14 because both g and $\partial g/\partial t$ are continuous on $[a,b]\times[-\tau,\tau]$, for some $\tau > 0$, where $g(x,t) = f(x, u(x)+th(x), u'(x)+th'(x))$.

Now we set this variation to zero for every $h \in X$. If we consider first those h which satisfy $h(a) = h(b) = 0$ we conclude that the integral term vanishes for all such h. Using the fundumental theorem (with $\Omega = (a,b)$) we see that the integrand is identically zero on $a < x < b$. This shows (4.2.2).

Now that we know that the integrand is zero, we can consider an $h \in X$ for which $h(b) = 0$ but $h(a) \neq 0$; we conclude that (4.2.3) holds. Similarly if $h(a) = 0$ but $h(b) \neq 0$; we conclude that (4.2.4) holds. □

4.2.2 Example (arc length)**.** Let's minimize the functional (4.0.2). The first variation is

$$\begin{aligned}
\frac{d}{dt}F(u+th)|_{t=0} &= \int_a^b \frac{d}{dt}\sqrt{1 + (u'(x) + th'(x))^2}|_{t=0}\,dx \\
&= \int_a^b \left[\frac{(u'(x) + th'(x))h'(x)}{\sqrt{1 + (u'(x) + th'(x))^2}}\right]_{t=0} dx = \int_a^b \left[\frac{u'(x)h'(x)}{\sqrt{1 + (u'(x))^2}}\right] dx \\
&= \left[\frac{u'(x)h(x)}{\sqrt{1 + (u'(x))^2}}\right]_{x=a}^{b} - \int_a^b \frac{d}{dx}\left[\frac{u'(x)}{(1 + u'(x)^2)^{1/2}}\right] h(x)\,dx \ .
\end{aligned}$$

Considering for the moment only those h for which $h(a) = h(b) = 0$ we conclude that the last integral must vanish for all such h. The Fundamental Theorem 4.1.10 implies

$$\frac{d}{dx}\left[\frac{u'(x)}{(1 + u'(x)^2)^{1/2}}\right] = 0$$

on $a < x < b$. This means $u' \equiv c$, a constant, on (a, b) so that u must be a straight line. If we now consider as well those h which do not vanish at $x = a$ but which still satisfy $h(b) = 0$ (and having already shown that the integral is zero) we see that

$$-\frac{u'(a)h(a)}{\sqrt{1+(u'(a))^2}} = 0$$

or that $u'(a) = 0$. Thus, u is a straight line with zero slope at $x = a$, i.e., $u \equiv$ constant. Finally, if we consider those h for which $h(b) \neq 0$ we similarly see that $u'(b) = 0$; but this gives us no new information. The best we can conclude is that u is a constant function and this is consistent with our geometric knowledge.

End Point Constraints We can optimize integrals of the form (4.2.1) subject to one or both of the contraints

$$u(a) \;=\; \alpha\;, \qquad\qquad u(b) \;=\; \beta \tag{4.2.5}$$

for some $\alpha, \beta \in \mathbb{R}$. In this case u satisfies the same Euler-Lagrange equation, but the conditions (4.2.5) replace the natural boundary conditions (4.2.3) and (4.2.4).

4.2.3 Corollary. *Assume the conditions of Theorem 4.2.1. Then an extremal $u \in C^2([a,b];\mathbb{R})$ of the functional F subject to the constraints (4.2.5) satisfies (4.2.2) but not, in general, (4.2.3) or (4.2.4).*

Proof. For this problem we make use of the corollary of Theorem 1.8.8 with the affine space Y equal to the set of all functions in $C^2([a,b];\mathbb{R})$ for which $v(a) = \alpha$ and $v(b) = \beta$. Then the variational direction h must be taken in the subspace X_0 equal to the set of all functions in $C^2([a,b];\mathbb{R})$ with $h(a) = 0$ and $h(b) = 0$. The calculation proceed as in the proof of the theorem, except the end point evaluations $[\,\frac{\partial f}{\partial u'}(x, u(x), u'(x))\,h(x)]_{x=a}^{x=b}$, from the integration by parts, is zero thanks to the constraint that $h \in X_0$ □

It is obvious from the proof of Theorem 4.2.1 that we can also impose only one of the two constraints in (4.2.5); for this problem the unconstrained end point will satisfy its natural boundary condition.

4.2.4 Example. Let's minimize the same functional $J(f)$ as in Example 4.2.2 except that we will require f to go through the endpoints $f(a) = \alpha$ and $f(b) = \beta$ for some $\alpha, \beta \in \mathbb{R}$. The computation of the variation is the same except that here the only valid variational 'directions', h, are those for which $h(a) = h(b) = 0$. Thus the first variation is

$$\frac{d}{dt}F(u+th)|_{t=0} \;=\; -\int_a^b \frac{d}{dx}\left[\frac{u'(x)}{(1+u'(x)^2)^{1/2}}\right] h(x)\,dx \;=\; 0\;.$$

We conclude that $u' \equiv c$, a constant, on (a, b). So we know that u is the equation of a line, but it need not be constant. Instead the constraints tell us it must be the line joining (a, α) and (b, β).

4.2.5 Example. Let $[a, b] \subset \mathbb{R}$ be a bounded interval and $\alpha, \beta \in \mathbb{R}$. If $u \in C^2([a, b])$ satisfies $u(a) = \alpha$ and $u(b) = \beta$ then 2π times the functional

$$F(u) \;=\; \int_a^b u(x)\,\sqrt{1+(u'(x))^2}\,dx$$

is the area of the surface of revolution created when the arc $(x, u(x))$, $a < x < b$, is rotated about the x-axis. We can find the Euler-Lagrange equation of the surface of revolution which has minimum surface area and spans the two circles of radius α and β at $x = a$ and $x = b$. Making use of the fact that the directions h of variation satisfy $h(a) = h(b) = 0$ we have

$$\begin{aligned}
\delta F(u;h) \;&=\; \int_a^b \frac{d}{dt}[(u+th)\sqrt{1+(u'+th')^2}]_{t=0}\,dx \\
&=\; \int_a^b \sqrt{1+(u')^2}\,h + \frac{u\,u'}{\sqrt{1+(u')^2}}\,h'\,dx \\
&=\; \int_a^b [\sqrt{1+(u')^2} - \frac{d}{dx}(\frac{u\,u'}{\sqrt{1+(u')^2}})]\,h\,dx\;.
\end{aligned}$$

The fundamental lemma then shows the Euler-Lagrange equations

$$\sqrt{1+(u')^2} - \frac{d}{dx}\left(\frac{u\,u'}{\sqrt{1+(u')^2}}\right) \;=\; 0$$

must hold on $a < x < b$.[26]

4.2.6 Exercise. Show that the Euler-Lagrange equation for the Brachistocrone of Example 4.1.3 is

...FILL THIS IN...

See Arfkin, p 780, for a solution. The solution curve is a cycloid; Arfkin gives a very interesting application of this to early pendulum clocks.[27]

4.3 Vector Valued Functions

There are at least three ways to generalize Theorem 4.2.1. We can consider optimizing integrals involving functions u of more than one variable, functions u which are vector valued, and integrands f which include higher order derivatives of u. The case when u is vector valued involves no new concepts, but it may be helpful to clarify notation.

If $f \in C^2([a,b]\times\mathbb{R}^{2r};\mathbb{R})$ and $u \in C^2([a,b];\mathbb{R}^r)$ with components $u_j(x)$, then

$$f(x,u,u') \;=\; f(x,u_1,...,u_r,u_1',...,u_r')$$

where we write $u = u(x)$, $u' = u'(x)$, etc. for short. We also use the notation

$$\frac{\partial f}{\partial u}(x,u,u') \;=\; \Big(\frac{\partial f}{\partial u_1}(x,u,u'),\cdots,\frac{\partial f}{\partial u_r}(x,u,u')\Big)$$

for the 'partial gradient' with respect to the vector variable u, and similarly for u'. Then the total derivative $\frac{df}{dx}(x,u(x),u'(x))$ with respect to x looks like

$$\frac{df}{dx} \;=\; \frac{\partial f}{\partial x} + \sum_{j=1}^{r}\frac{\partial f}{\partial u_j}\,\frac{du_j}{dx} + \sum_{j=1}^{r}\frac{\partial f}{\partial u_j'}\,\frac{du_j'}{dx}$$

where we have suppressed the arguments $(x,u,u') = (x,u(x),u'(x))$. We will sometimes shorten this last expression to

$$\frac{df}{dx} \;=\; \frac{\partial f}{\partial x} + \frac{\partial f}{\partial u}\cdot\frac{du}{dx} + \frac{\partial f}{\partial u'}\cdot\frac{du'}{dx}$$

where $\frac{\partial f}{\partial u}\cdot\frac{du}{dx}$ stands for the vector dot product between the partial gradient $\frac{\partial f}{\partial u}$ and $u' = \frac{du}{dx}$, and similarly for $\frac{\partial f}{\partial u'}\cdot\frac{du'}{dx} = \frac{\partial f}{\partial u'}\cdot u''$.

4.3.1 Theorem. *Assume $[a,b]\subset\mathbb{R}$ compact, $f\in C^2([a,b]\times\mathbb{R}^{2r};\mathbb{R})$, and set $X = C^2([a,b];\mathbb{R}^r)$. Let $F: X\to\mathbb{R}$ be defined by*

$$F(u) \;=\; \int_a^b f(x,u(x),u'(x))\,dx\;. \tag{4.3.1}$$

If $u\in X$ is a local extremal of F then u satisfies the vector differential equation

$$\frac{d}{dx}\Big(\frac{\partial f}{\partial u'}\Big) - \frac{\partial f}{\partial u} = 0 \tag{4.3.2}$$

[26] This differential equation can be solved explicitly by setting $v' = \sqrt{1+(u')^2}$ for some new function v. Then $(u\,u'/v')' = v$ from which $u\,u' = v\,v'$.... (STEVE: Finish this.) The solution is of the form $u(x) = \alpha\cosh(\frac{x-\beta}{\alpha})$ for constants of integration α and β. This function is called a catenary and its surface of revolution is a catenoid.

[27] TO STEVE: Arfkin points out alternative forms of the Euler-Lagrange equations; equation (17.16) he uses a lot!

or more explicitly

$$\frac{\partial^2 f}{\partial x\,\partial u_i'} + \sum_{j=1}^{r} \frac{\partial^2 f}{\partial u_j\,\partial u_i'}\,\frac{du_j}{dx} + \sum_{j=1}^{r} \frac{\partial^2 f}{\partial u_j'\,\partial u_i'}\,\frac{d^2u_j}{dx^2} - \frac{\partial f}{\partial u_i} \;=\; 0, \qquad i = 1, ..., r\ .$$

Furthermore, at the end points of the interval u satisfies

$$\frac{\partial f}{\partial u'}(a, u(a), u'(a)) = 0 \tag{4.3.3}$$

and

$$\frac{\partial f}{\partial u'}(b, u(b), u'(b)) = 0. \tag{4.3.4}$$

Proof. Using the first corollary of Theorem 4.1.10, the proof is only a vector version of the proof of Theorem 4.2.1, and we leave the details to the reader.[28] □

4.3.2 Exercise. State and prove a version of Corollary 4.2.3 that applies to the functional (4.3.1), i.e., when u is vector valued.

4.3.3 Example (Particle dynamics in a potential field)**.** To continue with example 4.1.2 let's find the Euler-Lagrange equations for the action integral

$$A(x) \;=\; \int_a^b \tfrac{1}{2}\dot{x}(t)^t M\dot{x}(t) - U(x(t), t)\,dt$$

where M is a positive definite mass matrix. Assuming that the beginning and ending locations of the system are given, $x(a)$ and $x(b)$ are fixed points in $\mathbb{R}^{3n}$ and the variational direction h satisfies $h(a) = 0$ and $h(b) = 0$ in $\mathbb{R}^{3n}$. Then we have

$$\delta A(x; h) \;=\; \int_a^b \dot{h}(t)^t M\dot{x}(t) - \nabla U(x(t), t)\,h(t)\,dt$$

where $\nabla U\,h = \sum_{j=1}^{3n} \frac{\partial U}{\partial x_j} h_j$. Integrating the kinetic energy term by parts, and using the zero boundary values of h, we obtain $M\ddot{x}(t) + \nabla U(x(t), t) = 0$ on $a < t < b$, if we interpret ∇U as a column vector.

4.4 Higher Order Derivatives

4.4.1 Theorem. *Assume $[a, b] \subset \mathbb{R}$, $f \in C^{m+1}([a, b] \times \mathbb{R}^{m+1}; \mathbb{R})$, and set $X = C^{2m}([a, b]; \mathbb{R})$. Let $F : X \to \mathbb{R}$ be defined by*

$$F(u) = \int_a^b f(x, u(x), u'(x), u''(x), ..., u^{(m)}(x))\ dx.$$

If $u \in X$ is a local extremal of F then u satisfies the Euler-Lagrange differential equation

$$\sum_{k=0}^{m} \Big(-\frac{d}{dx}\Big)^k \Big(\frac{\partial f}{\partial u^{(k)}}(x, u(x), u'(x), u''(x), ..., u^{(m)}(x))\Big) \;=\; 0\ . \tag{4.4.1}$$

Furthermore, at the end points of the interval u may also satisfy additional natural boundary conditions, depending on any constraints u is asked to satisfy at those points.

When $m = 2$ and u is not constrained at a or b, u will satisfy the natural boundary conditions

$$\begin{aligned} \frac{\partial f}{\partial u''}(a, u(a), u'(a), u''(a)) \;&=\; 0 \\ \frac{\partial f}{\partial u''}(b, u(b), u'(b), u''(b)) \;&=\; 0 \end{aligned}$$

[28] TO STEVE: Do the examples conform to the hypotheses of the theorems?

and

$$\begin{aligned}\frac{\partial f}{\partial u'}(a,u(a),u'(a),u''(a)) - \frac{d}{dx}\frac{\partial f}{\partial u''}(a,u(a),u'(a),u''(a)) &= 0\\ \frac{\partial f}{\partial u'}(b,u(b),u'(b),u''(b)) - \frac{d}{dx}\frac{\partial f}{\partial u''}(b,u(b),u'(b),u''(b)) &= 0\ ,\end{aligned}$$

as well as the Euler-Lagrange equation

$$\frac{\partial f}{\partial u} - \frac{d}{dx}\frac{\partial f}{\partial u'} + \frac{d^2}{dx^2}\frac{\partial f}{\partial u''} = 0$$

on $a < x < b$.

Proof. If u is an extremal then the first variation $\delta F(u;h) = 0$ for all $h \in X$ by Theorem 1.8.8. We calculate the variation, making use of our smoothness assumptions on f and u to justify differentiating under the integral, the chain rule, and integrate by parts.

$$\begin{aligned}\delta F(u;h) &= \frac{d}{dt}\left[\int_a^b f(x,u(x)+t\,h(x),u'(x)+t\,h'(x),...,u^{(m)}(x)+t\,h^{(m)}(x))\,dx\right]_{t=0}\\ &= \int_a^b \frac{d}{dt}\Big[f(x,u(x)+t\,h(x),u'(x)+t\,h'(x),...,u^{(m)}(x)+t\,h^{(m)}(x))\Big]_{t=0}\,dx\\ &= \int_a^b \Big[\sum_{k=0}^m \frac{\partial f}{\partial u^{(k)}}(x,u(x)+t\,h(x),...,u^{(m)}(x)+t\,h^{(m)}(x))\,h^{(k)}(x)\Big]_{t=0}\,dx\\ &= \sum_{k=0}^m \int_a^b \frac{\partial f}{\partial u^{(k)}}(x,u(x),...,u^{(m)}(x))\,h^{(k)}(x)\,dx\\ &= \sum_{k=0}^m\sum_{j=0}^k \Big[\Big(-\frac{d}{dx}\Big)^j\Big(\frac{\partial f}{\partial u^{(k)}}(x,u(x),u'(x),...,u^{(m)}(x))\Big)\,h^{(k-1-j)}(x)\Big]_{x=a}^{x=b}\\ &\qquad \sum_{k=0}^m \int_a^b \Big(-\frac{d}{dx}\Big)^k\Big(\frac{\partial f}{\partial u^{(k)}}(x,u(x),u'(x),...,u^{(m)}(x))\Big)\,h(x)\,dx\end{aligned}$$

where we have integrated by parts repeatedly. Now we set this variation to zero for every $h \in X$. If we consider first those h which satisfy $h^{(k)}(a) = h^{(k)}(b) = 0$, for $k = 0, 1, ..., m-1$, we conclude that the integral term vanishes for all such h. Using the fundumental theorem (with $\Omega = (a,b)$) we see that the integrand is identically zero on $a < x < b$. This shows (4.4.1).

Now that we know that the integrand is zero, we can consider $h \in X$ for which $h^{(k)}(a) = h^{(k)}(b) = 0$ except that $h^{(k)} \neq 0$ for one k and one end point; we conclude that the resulting natural boundary conditions hold. When $m = 2$ this yields the equations displayed in the theorem. □

4.4.2 Example (Cubic smoothing spline)**.** We carry out some analysis for Example 4.1.5. This example is a little more complicated, for care must be taken at the data locations x_j. It turns out that the minimizer u is not quite C^4, but only C^3, at those values of x. For concreteness we will assume $a < x_1 < \cdots < x_n < b$. The functional to be minimized is

$$F(u) = \int_a^b (u''(x))^2\,dx + \sum_{j=1}^n (y_j - u(x_j))^2$$

where, for simplicity, we have taken $\lambda = 1$. Setting $x_0 = a$ and $x_{n+1} = b$, the variation is

$$\begin{aligned}
\tfrac{1}{2}\delta F(u;h) &= \int_a^b [(u'' + th'')\, h'']_{t=0}\, dx + \sum_{j=1}^{n} [(y_j - u(x_j) - th(x_j))\,(-h(x_j))]_{t=0} \\
&= \sum_{j=0}^{n} \int_{x_j}^{x_{j+1}} u''\, h''\, dx - \sum_{j=1}^{n} (y_j - u(x_j))\, h(x_j) \\
&= \sum_{j=0}^{n} \Big\{ \int_{x_j}^{x_{j+1}} u^{(4)}\, h\, dx + \big[u''\, h'\big]_{x_j}^{x_{j+1}} - \big[u'''\, h\big]_{x_j}^{x_{j+1}} \Big\} - \sum_{j=1}^{n} (y_j - u(x_j))\, h(x_j)
\end{aligned}$$

where we have integrated by parts twice to get the last expression. By allowing h to be non-zero only on the open subintervals $(x_j, x_{j+1}) \subset (a, b)$ at first, we conclude from $\delta F(u;h) = 0$ that

$$u^{(4)} = 0\,, \qquad \text{on } x_j < x < x_{j+1}, \qquad j = 0, ..., n\ .$$

This true, we then allow $h(x_j) \neq 0$, and then $h'(x_j) \neq 0$, at one x_j at a time to conclude that

$$\begin{aligned}
u(x_j) - j_j + u'''(x_j+) - u'''(x_j-) &= 0\,, \qquad j = 1, ..., n \\
u''(a) = u''(b) = u'''(a) = u'''(b) &= 0 =
\end{aligned}$$

where x_j+ and x_j- are the right and left hand limits at x_j. The first equations are the 'knot conditions' of the spline, and the second are the natural end point boundary conditions.

4.4.3 Exercise. State and prove the analog of the preceding theorem, for the case $m = 2$, when u is $\mathbb{R}^r$ valued. Then, we must interpret $\partial f/\partial u^{(k)} = (\partial f/\partial u_1^{(k)}, ..., \partial f/\partial u_r^{(k)})$ as a gradient.

4.5 Several Independent Variables

Extension of the calculus of variations to problems of several variables requires the following multivariate fundumental theorem of calculus; a proof can be found in any text on advanced or multivariate calculus.

4.5.1 Theorem (divergence theorem)**.** *Let $\Omega \subset \mathbb{R}^d$ be a domain whose boundary is C^1 (???) with outward unit normal n at each point of $\partial\Omega$, and let $f \in C^1(\Omega; R^d)$. Then*

$$\int_\Omega \nabla \cdot f\, dx = \int_{\partial\Omega} n \cdot f\, d\sigma.$$

We can now prove a multivariate integration by parts formula.

4.5.2 Corollary. *Let $\Omega \subset \mathbb{R}^d$ be a domain whose boundary is C^1 (???) with outward unit normal n at each point of $\partial\Omega$, and let $f \in C^1(\Omega; R^d)$ be vector valued and $g \in C^1(\Omega; R)$ be scalar valued. Then*

$$\int_\Omega f \cdot \nabla g\, dx = -\int_\Omega g\, \nabla \cdot f\, dx + \int_{\partial\Omega} g\, n \cdot f\, d\sigma.$$

Proof. Begin with the product rule

$$\nabla \cdot (gf) = \nabla g \cdot f + g\, \nabla \cdot f.$$

Integrating this (scalar) equation over Ω gives

$$\int_\Omega \nabla g \cdot f\, dx = -\int_\Omega g\, \nabla \cdot f\, dx + \int_\Omega \nabla \cdot (gf)\, dx = -\int_\Omega g\, \nabla \cdot f\, dx + \int_{\partial\Omega} g\, n \cdot f\, d\sigma$$

where we have used the divergence theorem on the last term. □

The preceding formulas can be generalized to cases when g and f are higher order tensors or differential forms on a manifold Ω. The starting point is always some version of the product rule for differentiation followed by some version of the divergence (or Stoke's) theorem.[29] In this case it is sometimes helpful to include component indeces in the notation for a vector or tensor. For instance, a vector u would be denoted (u_i) or just u_i, and a second order tensor a by (a_{ij}) or just a_{ij}. With this notation we must always specify what the range of the ideces i and j are. For instance, $u = u(x)$ might be a $\mathbb{R}^r$ vector valued function of $x \in \mathbb{R}^d$; in this case the i in u_i runs from 1 to r. We know that if $A = (a_{ij})$ is a matrix and $u = (u_j)$ a vector, the vector $v = Au$ has components $v_i = \sum_j a_{ij}u_j$ provided the dimensions of A and u are comensurate. It is (a handy) tradition that this multiplication operation be shortened to $v_i = a_{ij}u_j$ where summation over any repeated index, j in this case, is implied without writing the summation sign explicitly.

A concise notation for differentiation of $u(x)$ with respect to the j-th independent variable x_j is

$$u_{i,j} \;=\; \frac{\partial u_i}{\partial x_j}, \qquad u_{i,jk} \;=\; \frac{\partial^2 u_i}{\partial x_j\,\partial x_k}, \qquad u_{ij,kl} \;=\; \frac{\partial^2 u_{ij}}{\partial x_k\,\partial x_\ell}, \qquad \text{etc.}$$

Suppose $\Omega \subset \mathbb{R}^d$ and $f \in C^1(\Omega;\mathbb{R}^d)$. In our new notation the divergence theorem looks like

$$\int_\Omega f_{i,i}(x)\,dx \;=\; \int_{\partial\Omega} n_i(x)\,f_i(x)\,ds(x)\;,$$

or just $\int_\Omega f_{i,i}\,dx = \int_{\partial\Omega} n_i f_i\,ds$, where the outward unit normal is n_i. If $g \in C^1(\Omega)$ the product rule looks like $(f_i g)_{,i} = f_{i,i}g + f_i g_{,i}$, and the integration by parts formula

$$\int_\Omega f_i\,g_{,i}\,dx \;=\; -\int_\Omega f_{i,i}\,g\,dx + \int_{\partial\Omega} n_i f_i\,g\,ds\;.$$

It is possible that f itself is a component of a higher order 'tensor' so that formulas like

$$\int_\Omega f_{ijk}\,g_{,k}\,dx \;=\; -\int_\Omega f_{ijk,k}\,g\,dx + \int_{\partial\Omega} n_k f_{ijk}\,g\,ds$$

hold as well.

Before we state the next theorem let's write down some notation for the various derivatives that will occur. For $x \in \Omega \subset \mathbb{R}^d$, $y \in \mathbb{R}$, and $z \in \mathbb{R}^d$ we will consider two functions, $u(x)$ and $f(x,y,z)$. We will set

$$\nabla u(x) \;=\; (\frac{\partial u}{\partial x_1}(x), ..., \frac{\partial u}{\partial x_d}(x))$$

and

$$\frac{\partial f}{\partial z} \;=\; (\frac{\partial f}{\partial z_1}, ..., \frac{\partial f}{\partial z_d})\;.$$

When we have the composition $f(x, u(x), \nabla u(x))$ we will usually write

$$\frac{\partial f}{\partial u} \;=\; \frac{\partial f}{\partial u}(x, u(x), \nabla u(x)) \;=\; \frac{\partial f}{\partial y}(x,y,z)\Big|_{(y,z)=(u(x),\nabla u(x))}$$

and

$$\frac{\partial f}{\partial(\nabla u)} \;=\; \frac{\partial f}{\partial(\nabla u)}(x, u(x), \nabla u(x)) \;=\; \frac{\partial f}{\partial z}(x,y,z)\Big|_{(y,z)=(u(x),\nabla u(x))}\;.$$

Now define $g \in C^1(\Omega)$ by $g(x) \;=\; f(x, u(x), \nabla u(x))$. Then we will write

$$\frac{df}{\partial x} \;=\; \frac{d}{\partial x} f(x, u(x), \nabla u(x)) \;=\; \nabla g(x)\;.$$

When g is the $\mathbb{R}^d$-valued function $g(x) = \partial f/\partial(\nabla u)(x, u(x), \nabla u(x))$ we also write

$$\frac{d}{\partial x}\cdot\frac{\partial f}{\partial(\nabla u)} \;=\; \frac{d}{\partial x}\cdot\frac{\partial f}{\partial(\nabla u)}(x, u(x), \nabla u(x)) \;=\; \nabla\cdot g(x)$$

[29] TO STEVE: I should include exercises on this.

for the divergence of g. If n is the unit outward normal vector on $\partial\Omega$ we will use

$$\frac{\partial u}{\partial n}(x) \;=\; n(x)\cdot\nabla u(x)$$

to denote the directional derivative of u in the direction of n at the point $x\in\partial\Omega$. *We will usually omit reference to the variables, e.g., x or $x, u(x), \nabla u(x)$, in the preceding formulas.*

4.5.3 Theorem. *Assume $\Omega\subset\mathbb{R}^d$ with boundary sufficiently smooth that the divergence theorem holds on Ω, let $f\in C^2(\bar\Omega\times\mathbb{R}^{1+d};\mathbb{R})$, and set $X=C^2(\bar\Omega;\mathbb{R})$. Let $F:X\to\mathbb{R}$ be defined by*

$$F(u) \;=\; \int_\Omega f(x,u(x),\nabla u(x))\,dx\ .$$

If $u\in X$ is a local extremal of F then u satisfies the Euler-Lagrange differential equation

$$\frac{d}{\partial x}\cdot\frac{\partial f}{\partial(\nabla u)}-\frac{\partial f}{\partial u} \;=\; 0\ . \tag{4.5.1}$$

Furthermore, for $x\in\partial\Omega$, u satisfies

$$n\cdot\frac{\partial f}{\partial(\nabla u)}(x,u(x),\nabla u(x)) \;=\; 0\ . \tag{4.5.2}$$

Proof. The logic of this proof is the same as for the case when $\Omega=(a,b)$; the calculation of the variation is slightly different. Again our smoothness assumptions on f and u allow us to integrate under the integral, apply the chain rule, and integrate by parts.

$$\begin{aligned}
\delta F(u;h) &= \frac{d}{dt}\left[\int_\Omega f(x,u(x)+t\,h(x),\nabla u(x)+t\,\nabla h(x))\,dx\right]_{t=0}\\
&= \int_\Omega \frac{d}{dt}\Big[f(x,u(x)+t\,h(x),\nabla u(x)+t\,\nabla h(x))\Big]_{t=0}\,dx\\
&= \int_\Omega\Big[\frac{\partial f}{\partial u}(x,u(x)+t\,h(x),\nabla u(x)+t\,\nabla h(x))\,h(x)\\
&\qquad\qquad +\frac{\partial f}{\partial(\nabla u)}(x,u(x)+t\,h(x),\nabla u(x)+t\,\nabla h(x))\cdot\nabla h(x)\Big]_{t=0}\,dx\\
&= \int_\Omega\Big[\frac{\partial f}{\partial u}(x,u(x),\nabla u(x))\,h(x)+\frac{\partial f}{\partial(\nabla u)}(x,u(x),\nabla u(x))\cdot\nabla h(x)\Big]\,dx\\
&= \int_\Omega\frac{\partial f}{\partial u}(x,u(x),\nabla u(x))\,h(x)\;dx+\int_{\partial\Omega}n\cdot\frac{\partial f}{\partial(\nabla u)}(x,u(x),\nabla u(x))\,h(x)\;d\sigma(x)\\
&\qquad\qquad -\int_\Omega\frac{d}{\partial x}\cdot\frac{\partial f}{\partial(\nabla u)}(x,u(x),\nabla u(x))\,h(x)\;dx\\
&= \int_\Omega\Big[\frac{\partial f}{\partial u}(x,u(x),\nabla u(x))-\frac{d}{\partial x}\cdot\frac{\partial f}{\partial(\nabla u)}(x,u(x),\nabla u(x))\Big]\,h(x)\,dx\\
&\qquad\qquad +\int_{\partial\Omega}n\cdot\frac{\partial f}{\partial(\nabla u)}(x,u(x),\nabla u(x))\,h(x)\;d\sigma(x)
\end{aligned}$$

Now we set this variation to zero for every $h\in X$. If we consider first those h which satisfy $h(x)=0$ on $\partial\Omega$ we conclude that the first integral vanishes for all such h. Using the fundumental theorem for the calculus of variations we see that the integrand is identically zero on Ω. This shows (4.5.1).

Since (4.5.1) is true we have

$$\delta F(u;h) \;=\; \int_{\partial\Omega}n\cdot\frac{\partial f}{\partial(\nabla u)}(x,u(x),\nabla u(x))\,h(x)\,d\sigma(x)$$

at an extremal u. Any function in $C^\infty(\mathbb{R}^d)$, when restricted to $\bar\Omega$, is a valid choice for h; and if $n\cdot(\partial f/\partial(\nabla u))$ is not zero at some point $x\in\partial\Omega$ then we may choose some smooth h which is zero on $\mathbb{R}^d$ except in a small neighborhood of x to get a contradition to the requirement $\delta F(u;h)=0$ for all h.[30] Thus (4.5.2) holds. □

[30] TO STEVE: This is an approach to the corollary of the fundamental lemma; I should tie the two together.

4.5.4 Example (Laplace's equation)**.** Consider the functional

$$J(u) = \int_\Omega |\nabla u(x)|^2 \, dx = \int_\Omega \sum_{j=1}^d (\partial_j u(x))^2 \, dx$$

defined for real valued $u \in H^1(\Omega)$. It is obvious that $J(u) \geq 0$ for all functions u whose second derivatives belong to $L^2(\Omega)$; and it is clear that this minimum is attained for all constant functions. In order to make the problem more interesting and useful we consider the affine subspace of $H^1(\Omega)$ given by

$$H_g = \{u \; ; \; u = g \text{ for } x \in \partial\Omega\}.$$

Here we let g be any continuous function on $\partial\Omega$, and we assume that $\partial\Omega$ is sufficiently smooth.[31] Since $u = g$ on $\partial\Omega$, and since $u(x) + th(x)$ must also lie in H_g for all $t \in \mathbb{R}$, the valid variations h for J are those which satisfy $h = 0$ on $\partial\Omega$. The first variation of J is

$$\begin{aligned}
\frac{d}{dt} J(u + th)\Big|_{t=0} &= \int_\Omega \sum_{j=1}^d \frac{d}{dt} (\partial_j u(x) + t\partial_j h(x))^2 \Big|_{t=0} \, dx \\
&= \sum_{j=1}^d \int_\Omega \frac{d}{dt} (\partial_j u(x)^2 + 2t\partial_j u(x)\partial_j h(x) + t^2 \partial_j h(x)^2) \Big|_{t=0} \, dx \\
&= 2 \int_\Omega \nabla u(x) \cdot \nabla h(x) \, dx \\
&= 2 \int_{\partial\Omega} h(x) \, \partial_n u(x) \, d\sigma(x) - 2 \int_\Omega h(x) \, \Delta u(x) \, dx
\end{aligned}$$

where we have applied Corollary 4.5.2 with $f = \nabla u$ and $g = h$. But since $h = 0$ on $\partial\Omega$ we see that

$$\delta J(u, h) = -2 \int_\Omega h(x) \, \Delta u(x) \, dx.$$

By the Fundumantal Theorem $\delta J(u, h) = 0$ implies

$$\Delta u(x) = 0 \qquad\qquad x \in \Omega.$$

The solution of this variational problem is thus also the solution of the Dirichlet problem: *Find* $u \in C(\bar{\Omega}) \cap H^1(\Omega)$ *such that* $\Delta u = 0$ *on* Ω *and* $u = g$ *on* $\partial\Omega$.

4.5.5 Example (Poisson's equation)**.** A useful extension of the last example is to minimize the functional

$$J(u) = \int_\Omega |\nabla u(x)|^2 \, dx + 2 \int_\Omega f(x) u(x) \, dx.$$

To the variation above we add the term $2 \int_\Omega f(x) h(x) \, dx$, which needs no integration by parts. The Euler-Lagrange equations become

$$\Delta u(x) = f(x) \qquad\qquad x \in \Omega.$$

The solution of this variational problem is thus also the solution of the Dirichlet problem: *Find* u *in a suitable space of functions such that* $\Delta u = f$ *on* Ω *and* $u = g$ *on* $\partial\Omega$.

4.5.6 Example (minimal surface equation)**.** Let $u \in C^2(\bar{\Omega})$ where $\Omega \subset \mathbb{R}^d$ is a bounded open set. The 'area' of the d-dimensional surface $\{(x, u(x)) \; ; \; x \in \Omega\}$ in $\mathbb{R}^{d+1}$ is given by

$$F(u) \; = \; \int_\Omega \sqrt{1 + |\nabla u(x)|^2} \, dx \; .$$

[31] TO STEVE: Technically we should be able to extend g to a $H^1(\Omega)$ function on all of Ω. I have to clear up some technicalities in this example.

We compute the variation of F:

$$\begin{aligned}\delta F(u;h) &= \int_\Omega \frac{d}{dt}\Big|_{t=0} \sqrt{1+|\nabla u(x)+t\,\nabla h(x)|^2}\,dx = \int_\Omega \frac{\nabla u\cdot\nabla h}{\sqrt{1+|\nabla u|^2}}\,dx \\ &= -\int_\Omega \nabla\cdot\Big(\frac{1}{\sqrt{1+|\nabla u|^2}}\nabla u\Big)\,h\,dx + \int_{\partial\Omega}\frac{1}{\sqrt{1+|\nabla u|^2}}\,n\cdot\nabla u)\,h\,ds\end{aligned}$$

where n is the outward unit normal on $\partial\Omega$. We conclude that (sufficiently smooth) surfaces of minimal area are graphs of functions satisfying

$$\nabla\cdot\Big(\frac{1}{\sqrt{1+|\nabla u|^2}}\nabla u\Big) = 0$$

in Ω. If the surface is constrained by specifying $u=g$ on $\partial\Omega$ for some $g\in C(\partial\Omega)$, then u must also satisfy this constraint. If u is unconstrained on $\partial\Omega$ then u also satisfies the natural boundary conditions $n\cdot\nabla u=0$ on $\partial\Omega$.

4.5.7 Example (wave equation)**.** We calculate the variation of the functional in Equation (4.1.2).

$$\begin{aligned}&\int_0^T\int_a^b \frac{d}{d\epsilon}\Big[\tfrac{1}{2}\rho\big(\frac{\partial u}{\partial t}+\epsilon\frac{\partial h}{\partial t}\big)^2-\tfrac{1}{2}\tau\big(\frac{\partial u}{\partial x}+\epsilon\frac{\partial h}{\partial x}\big)^2\Big]_{\epsilon=0}\,dx\,dt \\ &\quad= \int_0^T\int_a^b \rho\frac{\partial u}{\partial t}\frac{\partial h}{\partial t}-\tau\frac{\partial u}{\partial x}\frac{\partial h}{\partial x}\,dx\,dt \\ &\quad= \int_0^T\int_a^b -\rho\frac{\partial^2 u}{\partial t^2}\,h+\tau\frac{\partial^2 u}{\partial x^2}\,h\,dx\,dt+\int_a^b \rho\Big[\frac{\partial u}{\partial t}\,h\Big]_{t=0}^T dx-\int_0^T\tau\Big[\frac{\partial u}{\partial x}\,h\Big]_{x=a}^b dt\end{aligned}$$

where we have integrated by parts, the first integrand with respect to t and the second integrand with respect to x. If this expression is set to zero for any $h\in C_0^\infty([a,b]\times[0,T])$ we conclude that u, if smooth enough, must satisfy the wave equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\tau}{\rho}\frac{\partial^2 u}{\partial x^2}$$

on $(a,b)\times(0,T)$. Due to the rectangular shape of our domain we have not needed the divergence theorem to integrate by parts in this example.

If u is constrained at a and b ($u(a,t)=\alpha$ and $u(b,t)=\beta$ for constants α and β, for all $t\in[0,T]$ for instance) then we require $h(a,t)=0$ and $h(b,t)=0$ for all $t\in[0,T]$. In this case u sataifies these boundary conditions as well as the wave equation. If u is unconstrained at the end points $x=a$ and $x=b$ then we must allow h to be non-zero at $x=a$ and $x=b$ when $0<t<T$. This implies that u must also satisfy the natural boundary conditions

$$\frac{\partial u}{\partial x}(a,t) = \frac{\partial u}{\partial x}(b,t) = 0$$

for all $t\in(0,T)$.

4.6 Vector Functions of Several Variables

We now allow the function u of Theorem 4.5.3 to be vector valued. We will make use of the index notation introduced in section 4.5.

4.6.1 Theorem. *Assume $\Omega\subset\mathbb{R}^d$ with boundary sufficiently smooth that the divergence theorem holds on Ω, let $f\in C^2(\bar\Omega\times\mathbb{R}^{r+rd};\mathbb{R})$, and set $X=C^2(\bar\Omega;\mathbb{R}^r)$. Let $F:X\to\mathbb{R}$ be defined by*

$$F(u) = \int_\Omega f(x,u(x),\nabla u(x))\,dx$$

where ∇u is the rd array with components $u_{j,i}=\partial u_j/\partial x_i$. If $u\in X$ is a local extremal of F then u satisfies the Euler-Lagrange system of differential equations

$$\frac{\partial f}{\partial u_i}-\sum_{j=1}^d\frac{d}{dx_j}\Big(\frac{\partial f}{\partial u_{i,j}}\Big) = 0\,,\qquad i=1,\ldots,r\,, \tag{4.6.1}$$

on Ω. Furthermore, on $\partial\Omega$, u satisfies

$$\sum_{j=1}^{d} n_j \frac{\partial f}{\partial u_{i,j}} \;=\; 0 \;, \qquad\qquad i = 1, ..., r \;. \tag{4.6.2}$$

Proof. We will only exhibit the calculation of the variation using index notation; summation over repeated indices is understood. As before $f = f(x, u(x), \nabla u(x)) = f(x, ..., u_i, ..., u_{i,j}, ...)$.

$$\begin{aligned}
\delta F(u;h) \;&=\; \int_\Omega \frac{d}{dt}\Big[f(x, u(x) + t\,h(x), \nabla u(x) + t\,\nabla h(x))\Big]_{t=0}\, dx \\
&=\; \int_\Omega \frac{\partial f}{\partial u_i}\, h_i + \frac{\partial f}{\partial u_{i,j}}\, h_{i,j}\, dx \\
&=\; \int_\Omega \frac{\partial f}{\partial u_i}\, h_i - (\frac{\partial f}{\partial u_{i,j}})_{,j}\, h_i\, dx + \int_{\partial\Omega} n_j\, \frac{\partial f}{\partial u_{i,j}}\, h_i\, d\sigma \;.
\end{aligned}$$

Here we have applied the integration by parts formula

$$\int_\Omega \frac{\partial f}{\partial u_{i,j}}\, h_{i,j}\, dx \;=\; \int_{\partial\Omega} n_j\, \frac{\partial f}{\partial u_{i,j}}\, h_i\, d\sigma - \int_\Omega (\frac{\partial f}{\partial u_{i,j}})_{,j}\, h_i\, dx$$

which is derived from the product rule

$$(\frac{\partial f}{\partial u_{i,j}}\, h_i)_{,j} \;=\; \frac{\partial f}{\partial u_{i,j}}\, h_{i,j} + (\frac{\partial f}{\partial u_{i,j}})_{,j}\, h_i$$

and the divergence theorem. Notice that

$$(\frac{\partial f}{\partial u_{i,j}})_{,j} \;=\; \frac{d}{\partial x_j}\frac{\partial f}{\partial u_{i,j}} \;.$$

□

We can also consider functionals which depend on higher order derivatives of u.[32]

4.6.2 Theorem. *Assume $\Omega \subset \mathbb{R}^d$ with boundary sufficiently smooth that the divergence theorem holds on Ω, let $f \in C^3(\bar\Omega \times \mathbb{R}^{r+rd+rd^2}; \mathbb{R})$, and set $X = C^4(\bar\Omega; \mathbb{R}^r)$. Let $F : X \to \mathbb{R}$ be defined by*

$$F(u) \;=\; \int_\Omega f(x, ..., u_i, ..., u_{i,j}, ..., u_{i,jk}, ...)\, dx$$

where $u_{i,j} = \partial u_i/\partial x_j$ and $u_{i,jk} = \partial^2 u_i/(\partial x_j\, \partial x_k)$ for $i = 1, ..., r,\ j, k = 1, ..., d$. If $u \in X$ is a local extremal of F then u satisfies the Euler-Lagrange system of differential equations

$$\frac{\partial f}{\partial u_i} - \sum_{j=1}^{d} \frac{d}{\partial x_j}(\frac{\partial f}{\partial u_{i,j}}) + \sum_{j=1}^{d}\sum_{k=1}^{d} \frac{d^2}{\partial x_j\, \partial x_k}(\frac{\partial f}{\partial u_{i,jk}}) \;=\; 0 \;, \qquad i = 1, ..., r \;, \tag{4.6.3}$$

on Ω. Furthermore, on $\partial\Omega$, u satisfies[33]

$$\sum_{j=1}^{d} n_j \frac{\partial f}{\partial u_{i,j}} \;+ ... = \; 0 \;, \qquad\qquad i = 1, ..., r \;. \tag{4.6.4}$$

[32] TO STEVE: I need to check details of this formulation; integration by parts twice here, leads to partial derivatives of h on $\partial\Omega$.

[33] TO STEVE: Correct this, and finish the proof.

Proof. We will only exhibit the calculation of the variation using index notation; summation over repeated indices is understood. As before $f = f(x, ..., u_i, ..., u_{i,j}, ..., u_{i,jk}, ...)$, and we keep in mind the integration by parts formula used in the proof of Theorem 4.6.1; we will need to apply it twice to get

$$\begin{aligned}\int_\Omega \frac{\partial f}{\partial u_{i,jk}}\, h_{i,jk}\, dx &= -\int_\Omega (\frac{\partial f}{\partial u_{i,jk}})_{,k}\, h_{i,j}\, dx + \int_{\partial\Omega} n_k \frac{\partial f}{\partial u_{i,jk}}\, h_{i,j}\, d\sigma \\ &= \int_\Omega (\frac{\partial f}{\partial u_{i,jk}})_{,jk}\, h_i\, dx - \int_{\partial\Omega} n_j\, (\frac{\partial f}{\partial u_{i,jk}})_{,k}\, h_i\, d\sigma + \int_{\partial\Omega} n_k \frac{\partial f}{\partial u_{i,jk}}\, h_{i,j}\, d\sigma\end{aligned}$$

Then,

$$\begin{aligned}\delta F(u;h) &= \int_\Omega \frac{\partial f}{\partial u_i}\, h_i + \frac{\partial f}{\partial u_{i,j}}\, h_{i,j} + \frac{\partial f}{\partial u_{i,jk}}\, h_{i,jk}\, dx \\ &= \int_\Omega \frac{\partial f}{\partial u_i}\, h_i - (\frac{\partial f}{\partial u_{i,j}})_{,j}\, h_i + (\frac{\partial f}{\partial u_{i,jk}})_{,jk}\, h_i\, dx \\ &\quad - \int_{\partial\Omega} n_j\, (\frac{\partial f}{\partial u_{i,jk}})_{,k}\, h_i\, d\sigma + \int_{\partial\Omega} n_k \frac{\partial f}{\partial u_{i,jk}}\, h_{i,j}\, d\sigma + \int_{\partial\Omega} n_j\, \frac{\partial f}{\partial u_{i,j}}\, h_i\, d\sigma\ .\end{aligned}$$

By first considering only those h which vanish near $\partial\Omega$ we conclude that (4.6.3) is true. □

4.6.3 Example (equations of solid mechanics[34])**.** Assume existence of a force density $w(n, x, t)$ at a point x in the body, at time t, and with orientation determined from a unit vector n. This force density is the vector function defined as follows. Let S be a (sufficiently smooth) surface with $x \in S$ and let n be a unit normal to S. For any (sufficiently small) $\epsilon > 0$ let $V_\epsilon^+ = V_\epsilon^+(S, x, t)$ be the material in the body which is inside the ball $B(x, \epsilon)$ and which is on the same side of S that n points to. And let $V_\epsilon^- = V_\epsilon^-(S, x, t)$ be the material in the body which is inside the ball $B(x, \epsilon)$ and which is on the same side of S that $-n$ points to. The material in V_ϵ^+ exerts a contact force upon the material in V_ϵ^-, perhaps called $w(S, \epsilon, x, t)$, across the surface S. As $\epsilon \downarrow 0$, this force vector has a limiting value $w(n, x, t) \in \mathbb{R}^d$ which depends on S only through its unit normal n. A key assumption of the continuum mechanical theory is that w depends linearly on n, that is, that there is a $T = T(x, t) \in \mathbb{R}^{d\times d}$ such that $w(n, x, t) = T(x, t)n$ for all $x \in \Omega$, times t, and unit vectors n in $\mathbb{R}^d$. The functions w and T are geometric objects, i.e., tensors, which are independent of the coordinate system used.

I WILL DERIVE this below: The equations (Gurtin, theorem p 101) of motion for the body Ω are $\nabla\cdot T + b = \rho\ddot{u}$, where $b = b(x, t)$ is the body force on Ω at x and t (gravitational or magnetic fields, for instance, that act on the body material from outside the body), $u = u(x, t)$ is the displacement vector of the body at x and t, and $\rho = \rho(x)$ is the mass density of the body material at x. Here the divergence of T is the vector whose i-th component (in rectangular coordinates) is $T_{ij,j} = \sum_{j=1}^d \partial T_{ij}/\partial x_j$, and $\ddot{u}$ denotes the second partial derivative of u with respect to time.[35]

In the small displacement theory of continuum mechanics $u(x, t)$, with $x \in \Omega \subset \mathbb{R}^d$, is the $\mathbb{R}^d$-valued displacement of a body Ω at the point x and at time t. The displacements are assumed sufficiently small that the force density at any point x in the body is a linear function of (only) the first order derivatives of u.[36] This is the stress tensor which ... CONSULT GURTIN ch X, on Linear Elasticity. The *strain* is the symmetric part of the gradient, ∇u,with respect to x, i.e., $\hat{\nabla} u = \frac{1}{2}(\nabla u + \nabla u^{\,t})$, where ^{t} denotes transpose of the square array (2nd order tensor) $u_{i,j} = \partial u_i/\partial x_j$.[37] There is then a linear transformation $\mathbf{C}$, the stiffness (4th order) tensor[38], which takes the strain to the stress (2nd order) tensor, $\mathrm{T} = \mathbf{C}[\hat{\nabla} u]$.

The total kinetic energy at time t of this system is $\frac{1}{2}\int_\Omega |\frac{\partial u}{\partial t}|^2\, dx$. And the total potential energy at time t is $\frac{1}{2}\int_\Omega \mathbf{C}[\hat{\nabla} u]\cdot\hat{\nabla} u\, dx$ where $A\cdot B = \sum_{i,j=1}^d a_{ij}b_{ij}$ denotes the matrix inner product for the two linear

[34] TO STEVE: Fix up this example.

[35] See Gurtin, *An Introduction to Continuum Mechanics*, for a treatment of the theory.

[36] Without the small displacement assumption one is led to the Navier-Stokes fluid dynamical equations.

[37] We are using rectangular coordinates here. To obtain the correct equations in other coordinate systems the tensor transformation rules can be applied. See the Appendix in Frankel's, *The Geometry of Physics*, or ANOTHER REFERENCE.

[38] It seems the letter **C** stands for compliance, which might be the inverse of stiffness, but the letter S is already in common use in the literature for both stress and strain.

transformations. We can find the dynamical equations of motion of this system by finding the functions $u \in C^2(\bar{\Omega} \times [0,\infty); \mathbb{R}^d)$ which make the functional

$$F(u) \;=\; \tfrac{1}{2}\int_\Omega |\partial_t^2 u|^2 - \mathbf{C}[\hat{\nabla} u]\cdot\hat{\nabla} u \; dx$$

stationary. For simplicity I want to use ∇u in place of $\hat{\nabla} u$; $\mathbf{C}$ has symmetry properties that make this OK, I think.

$$F(u) \;=\; \tfrac{1}{2}\int_\Omega \big[\frac{\partial u}{\partial t}\big]^2 - \mathbf{C}[\nabla u]\cdot\nabla u \; dx \; .$$

Using the linearity and symmetry of $\mathbf{C}$, $\mathbf{C}[\nabla u]\cdot\nabla h = \mathbf{C}[\nabla h]\cdot\nabla u$, the first variation of F is

$$\begin{aligned}
\delta F(u;h) \;&=\; \tfrac{1}{2}\int_\Omega \frac{d}{d\epsilon}\Big[\big[\frac{\partial u}{\partial t} + \epsilon\frac{\partial h}{\partial t}\big]^2 - \mathbf{C}[\nabla u + \epsilon\nabla h]\cdot[\nabla u + \epsilon\nabla h]\Big]_{\epsilon=0} \; dx \\
&=\; \int_\Omega \frac{\partial u}{\partial t}\cdot\frac{\partial h}{\partial t} - \mathbf{C}[\nabla u]\cdot\nabla h \; dx \\
&=\; \int_\Omega -\frac{\partial^2 u}{\partial t^2}\cdot h + \nabla\cdot\mathbf{C}[\nabla u]\cdot h \; dx - \int_{\partial\Omega} n\cdot\mathbf{C}[\nabla u]\cdot h \; ds
\end{aligned}$$

where we have integrated by parts in both t and x, where the divergence theorem was used.

If this variation is set to zero for all $h \in C_0^\infty(\bar{\Omega} \times [0,\infty); \mathbb{R}^d)$ we conclude that u must satisfy the wave equation

$$\frac{\partial^2 u}{\partial t^2} \;=\; \nabla\cdot\mathbf{C}[\nabla u]$$

for $x \in \Omega$. Depending on other boundary conditions we want to impose on u, it may also satisfy the natural boundary conditions

$$n\cdot\mathbf{C}[\nabla u] \;=\; 0$$

of zero normal stress on $\partial\Omega$.

4.7 Expansions in Orthogonal Functions

Certain optimization problems can be expressed so that they are reduced to optimizing individual component parts in a vector space basis. This is especially useful for optimizing a Hilbert space norm when an orthonormal basis is available.

To illustrate in the simplest case let's look again at the least squares problem in $\mathbb{R}^n$. We want to minimize $|X\beta - y|^2$ where $y \in \mathbb{R}^n$ is a column vector, $X = [x_1, ..., x_m]$ is $n \times m$ with $m < n$ and each $x_j \in \mathbb{R}^n$ is a column vector. The linear combination $X\beta = x_1\beta_1 + \cdots x_m\beta_m$ is supposed to be a good approximation to y. Suppose that the columns of X are orthonormal, so that $x_i' x_j = \delta_{ij}$ or $X'X = I$, the $m \times m$ identity matrix. Then

$$|X\beta - y|^2 = (\beta' X' - y')(X\beta - y) = \beta' X' X\beta - 2y' X\beta + y'y = |y|^2 + \sum_{j=1}^m (\beta_j^2 - 2y' x_j \beta_j).$$

Now, the valuable thing about this expression is that it is minimized by minimizing each term $\beta_j^2 - 2y' x_j \beta_j$ separately. This is because the variable in each term of the sum is independent of the variables in all other terms. Differentiating with respect to β_j gives $\beta_j = x_j' y$ which is the same as the normal equations $X'X\beta = X'y$ when $X'X = I$. In fact this diagonal decomposition of the normal equations is no more complicated if the x_j's are orthogonal but not necessarily of unit length. We then obtain $|x_j|^2\beta_j = x_j' y$, trivial to solve.

The following example in infinite dimensions is hardly more complicated once we know Parseval's theorem.

4.7.1 Example. Let $f \in L^2(0,\pi)$ be a given real-valued function, and consider the problem of finding a real function u which minimizes the integral

$$\int_0^\pi |u''(x) - f(x)|^2 \, dx \, ,$$

the L^2 distance of the second derivative u'' to f, subject to the boundary conditions $u(0) = u(\pi) = 0$. The set of finctions $\{c_n \sin nx \ ; \ n \in \mathbb{N}\}$ is a complete orthonormal set for $L^2(0,\pi)$ if $c_n = (\int_0^\pi \sin^2 nx\, dx)^{-1/2}$. And each function in this basis equals zero at 0 and π.

Now let's expand both f and u in a Fourier sine series:

$$u(x) = \sum_1^\infty a_n c_n \sin nx \qquad \text{and} \qquad f(x) = \sum_1^\infty b_n c_n \sin nx$$

where the b_n's are known and the a_n's unknown. (The converge here is in $L^2(0,\pi)$.) If $u'' \in L^2(0,\pi)$, $u''(x) = \sum_1^\infty -n^2 a_n \sin nx$ (also convergent in $L^2(0,\pi)$) and we see that

$$u''(x) - f(x) \ = \ \sum_1^\infty (-n^2 a_n - b_n)\, c_n \sin nx\,.$$

By Parseval's theorem

$$\int_0^\pi |u''(x) - f(x)|^2\, dx \ = \ \sum_{n=1}^\infty |-n^2 a_n - b_n|^2\,.$$

Now we see the advantage of this expression; it can be minimized by choosing, for each $n \in \mathbb{N}$, a_n to minimize $(-n^2 a_n - b_n)^2$, or $a_n = -b_n/n^2$. This u of course solves the equation $u'' = f$.

4.7.2 Example. This example comes from Courant and Hilbert, vol 1 pg 178. It asks us to minimize kenitic energy of a deformed surface (e.g., a square drum) subject to a specific value of the potential energy (I think!).

We want to minimize the functional

$$F(u) \ = \ \int_0^a \int_0^b (u_x^2 + u_y^2)\, dxdy$$

subject to the constraint

$$G(u) \ = \ \int_0^a \int_0^b u^2\, dxdy \ = \ 1\,.$$

Here $u = u(x,y)$ is real-valued and subscripts denote partial derivatives. In addition, we ask that $u = 0$ on the boundary of the square $0 \le x \le a, 0 \le y \le b$.

Due to the boundary conditions and the L^2 norms, we suspect that an expansion in the complete orthogonal (but not normalized) set $\sin(m\pi x/a)\, \sin(n\pi y/b)$ might be useful. Set

$$u(x,y) = \sum_{m=1}^\infty \sum_{n=1}^\infty c_{mn} \sin(m\pi x/a)\, \sin(n\pi y/b)\,.$$

Because the derivatives $\frac{d}{dx} \sin(m\pi x/a) = \frac{m\pi}{a} \cos(m\pi x/a)$ are also orthogonal, we obtain

$$F(u) \ = \ \sum_{m=1}^\infty \sum_{n=1}^\infty c_{mn}^2 (\frac{m^2}{a^2} + \frac{n^2}{b^2}) \qquad \text{and} \qquad G(u) \ = \ \frac{ab}{4} \sum_{m=1}^\infty \sum_{n=1}^\infty c_{mn}^2\,.$$

The numbers $(m^2/a^2 + n^2/b^2)$ grow with the indices m and n so $F(u)$ is minimized by putting as much weight as possible on the c_{mn}'s with lower indices. And this can be accomplished by setting every $c_{mn} = 0$ except c_{11}. The constraint $G(u) = 1$ implies that $c_{11} = 2/\sqrt{ab}$. Therefore the minimizing function is

$$u(x,y) \ = \ \frac{2}{\sqrt{ab}} \sin(\pi x/a)\, \sin(\pi y/b)\,.$$

4.7.3 Example. I would like to illustrate the Fredholm alternative by minimizing

$$\int_0^\pi |u''(x) - f(x)|^2\, dx\,,$$

with Neumann boundary conditions. Use cosine series on $(0,\pi)$...

4.7.4 Exercise. Let $f \in L^2(0,\pi)$ be given. Find a function u which minimizes the integral

$$\int_0^\pi |au(x) - bu''(x) - f(x)|^2 \, dx \,,$$

where a and b are positive constants. Use the complete orthonormal set $\{c_n \sin nx \ ; \ n \in \mathbb{N}\}$ as in a previous example.

4.8 More Applications and Examples

One example should be a spline fit to continuous data: minimize over all $u \in C^2$ or C^4 the functional

$$J(u) \ = \ \int_0^1 |g(x) - u(x)|^2 + |u''(x)|^2 \, dx$$

where g is a given function in $L^2(0,1)$. Notice that without the smoothing term $|u''|$ this problem would have no solution since C^2 is dense in L^2.

4.8.1 Example (Smoothing splines in d variables)**.** Let $\Omega \subset \mathbb{R}^d$ be open and $\xi_1, ..., \xi_n$ be distinct points in Ω, and let $y_1, ..., y_n$ be any real numbers (observations). The smoothing spline u which approximates the data y_j at ξ_j is the minimizer of

$$F(v) \ = \ \int_\Omega |\Delta^k u(x)|^2 \, dx + \lambda \sum_{j=1}^n |y_j - v(\xi_j)|^2 \ .$$

Here λ is a parameter that determines the trade-off between fidelity to the data values y_j and smoothness of u, and $k \in \mathbb{N}$ is the degree of smoothing imposed on u.

4.8.2 Example (Optimal kernel density estimator)**.**

4.8.3 Example (Opinion pooling)**.** We follow the references: "Generalized Opinion Pooling" by A. Garg, et al,

`http://rutcor.rutgers.edu/~amai/aimath04/AcceptedPapers/Garg-aimath04.pdf ,`

and section 33 of *Information Theory, Inference and Learning Algorithms* by David MacKay, Cambridge University Press, 2003.

In this context an 'opinion' is a probability density that reflects someone's notion of how values of a sample space should be weighted. (This person is often called an 'expert.') The goal of opinion pooling is to combine the opinions of many experts into a single opinion that is useful to someone who needs this information to make decisions. For instance, in business the experts are often the customers and a company wants to know what its customers think of a certain product. If that product is described by a tuple of variables, $x = (x_1, ..., x_d)$, the opinion of the i-th expert is a probability density (continuous, discrete, or mixed) $p_i(x)$. The product might be a car, and the variables consumers have opinions about might include performance, safety, gas mileage, price, etc. In order to design next year's model, company executives must understand what customers as a whole think of, or want in, the vehicle.

Let's assume that the tuple of random variables x takes values in an open subset Ω of $\mathbb{R}^d$ and that we have n (perhaps a large number) opinions $p_1(x), ..., p_n(x)$ which we wish to combine. In order to apply the variational methods in infinite dimensional vector spaces we will assume these densities represent continuous random variables. If we measure the 'distance' between any two probability densities by the L^2 norm

$$\Delta(f,g) \ = \ \|f - g\|_{L^2} \ = \ \Big(\int_\Omega |f(x) - g(x)|^2 \, dx \Big)^{1/2} \,,$$

we might choose as the pooled density the function $q(x)$ that minimizes the sum of squares of the distances to each p_i:

$$F(q) \ = \ \sum_{i=1}^n \|p_i - q\|_{L^2}^2 \ = \ \sum_{i=1}^n \int_\Omega |p_i(x) - q(x)|^2 \, dx$$

assuming all such integrals are finite. In fact we want to minimize $F(q)$ subject to the constraint $\int q(x)\,dx = 1$. We also require that $q(x) \geq 0$ for all $x \in \Omega$ but we will see that this will be satisfied without any special care. In fact for this particular Δ even the constraint that q integrate to 1 will automatically be satisfied.

The first variation of F is

$$\delta F(q;h) \;=\; \frac{d}{dt}\Big|_{t=0} \sum_{i=1}^{n} \int_\Omega |p_i(x) - q(x) - th(x)|^2\,dx \;=\; -2\int_\Omega \sum_{i=1}^{n} (p_i(x) - q(x))h(x)\,dx\ .$$

Setting this equal to zero for all h shows that the integrand $\sum_{i=1}^n (p_i(x) - q(x))$ must be the zero function on Ω. So the pooled opinion is

$$q(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} p_i(x)\ ,$$

the average of all the expert opinions.

There are several other common choices for the comparison function Δ. The 'relative entropy between the probability density functions p(x) and q(x)' on Ω is

$$\Delta(p,q) \;=\; \int_\Omega p(x)\,\log(\frac{p(x)}{q(x)})\,dx\ .$$

Since $\Delta(p,q) \neq \Delta(q,p)$ this expression does not define a metric on a vector space of functions on Ω; nonetheless it is a useful measure in information theory. Let's consider a similar pooling procedure by choosing q to minimize

$$F(q) \;=\; \sum_{i=1}^{n} \int_\Omega p_i(x)\,\log(\frac{p_i(x)}{q(x)})\,dx$$

subject to the constraint that

$$G(q) \;=\; \int_\Omega q(x)\,dx - 1 \;=\; 0\ .$$

Introducing a Lagrange multiplier we calculate the variation

$$\begin{aligned}\delta[F+\lambda G](q;h) \;&=\; \frac{d}{dt}\Big|_{t=0} \int_\Omega \sum_{i=1}^{n} p_i(x)\,[\log p_i(x) - \log(q(x)+th(x))] \;+\; \lambda[q(x)+th(x)]\;dx \\ &=\; \int \big[-\sum \frac{p_i(x)}{q(x)} + \lambda\big]\,h(x)\,dx\ .\end{aligned}$$

We conclude that $\sum \frac{p_i(x)}{q(x)} = \lambda$ on Ω, or that $q(x) = \lambda^{-1}\sum p_i(x)$ if $\lambda \neq 0$. The constraint equation shows that $\lambda^{-1}\int \sum p_i(x)\,dx = 1$, so $\lambda = n$ if $\int p_i(x)\,dx = 1$ for all i.

4.8.4 Exercise (Opinion pooling)**.** Let Ω be an open subset in $\mathbb{R}^d$ and $\Delta(q,p) = \int_\Omega q(x)\,\log(\frac{q(x)}{p(x)})\,dx$ be the relative entropy between the densities q and p on Ω. Let $F(q) = \sum_{i=1}^n \Delta(q,p_i)$. Show that the minimum of F, subject to the constraint $\int_\Omega q(x)\,dx = 1$, is

$$q(x) \;=\; p(x)/\int p(x)\,dx$$

where $p(x) = [\prod_1^n p_i(x)]^{1/n}$.

4.8.5 Example (Bayesian parameter estimate)**.** Without explaining the background in Bayesian decision theory or statistics:

Let $p(\theta)$ be the prior probability density for a parameter $\theta \in \Theta$ a subset of $\mathbb{R}^d$. And let $p(x|\theta)$ be the probability density for the variable $x \in \Omega$ conditioned on the unknown parameter θ. Here Ω is a subset of $\mathbb{R}^n$; we assume Ω has non-empty interior and that x is a 'continuous' random variable. The joint density of x and

θ is then $p(x,\theta) = p(x|\theta)\,p(\theta)$, and the marginal of x alone is $p(x) = \int_\Theta p(x|\theta)\,p(\theta)\,d\theta$. And finally using Bayes rule the conditional density of θ given x is

$$p(\theta|x) \;=\; \frac{p(x,\theta)}{p(x)} \;=\; \frac{p(x|\theta)\,p(\theta)}{\int_\Theta p(x|\theta)\,p(\theta)\,d\theta}\;.$$

In applications, $p(x|\theta)$ and $p(\theta)$ are usually known (at least by an educated guess), and then $p(x,\theta)$, $p(x)$ and $p(\theta|x)$ can be computed. For instance we might assume that our data is approximately normally distributed but not know the mean or variance. Then

$$p(x,\theta) \;=\; p(x|\mu,\sigma) \;=\; \frac{1}{\sqrt{2\pi}\sigma}\,\exp\big[-\frac{1}{2}(\frac{x-\mu}{\sigma})\big]\;.$$

Generally we will have a random sample from this distribution so the density $p(x_1,...,x_n|\mu,\sigma)$ is a product of the case for one x with the same unknown parameters. We also need to postulate a model (probability density) $p(\mu,\theta)$ for the unknown parameters to be able to find the densities needed to perform Bayesian estimation of μ and σ.

The objective of Bayesian estimation is to use the value of x (a random vector before the experiment but a vector of numbers, the data, after the experiment is done) to make a good estimate of the unknown parameter θ. That is, we want a rule, a function, that returns an estimate of $\theta \in \Theta$ whenever we give it some data $x \in \Omega$. The function that takes the data x to a parameter estimate θ is called the decision function; we denote it by $u : \Omega \to \Theta$. We would like to choose the best decision function in the sense that u minimizes the 'loss' $\lambda(\theta, u(x))$ when θ and x are given. The function $\lambda : \Theta \times \Omega \to \mathbb{R}$ is real valued and analogous to a metric. For instance a *quadratic* or *squared error loss* function is

$$\lambda(\theta, u(x)) \;=\; |\theta - u(x)|^2$$

where $|\cdot|$ is the Euclidean distance in $\mathbb{R}^d$.

But of course this loss depends on the unknown θ and x, also unknown before the experiment, and this is not a well posed optimization problem. Instead we choose u to minimize the *risk* or *expected loss*

$$R(u) \;=\; \int_\Omega \int_\Theta \lambda(\theta, u(x))\,p(x,\theta)\,d\theta\,dx \;=\; \int_\Omega \int_\Theta \lambda(\theta, u(x))\,p(\theta|x)\,d\theta\;p(x)\,dx$$

where we have computed $p(\theta|x)$ and $p(x)$.

To minimize R with respect to u we calculate the first variation in the direction v. To simplify the discussion we take the useful special case $\lambda(\theta, u(x)) = |\theta - u(x)|^2$. The variation is then

$$\begin{aligned}
\delta R(u;v) \;=\; \frac{d}{dt}\Big|_{t=0} R(u+tv) \;&=\; \int_\Omega \int_\Theta \frac{d}{dt}\Big|_{t=0} |\theta - u(x) - tv(x)|^2\,p(\theta|x)\,d\theta\;p(x)\,dx \\
&=\; -2\int_\Omega \int_\Theta (\theta - u(x))\cdot v(x)\,p(\theta|x)\,d\theta\;p(x)\,dx \\
&=\; -2\int_\Omega \Big[\int_\Theta \theta\,p(\theta|x)\,d\theta - u(x)\Big]\cdot v(x)\;p(x)\,dx
\end{aligned}$$

where $\cdot$ denotes the dot product in $\mathbb{R}^d$. Since v is an arbitrary vector function we conclude that

$$u(x) \;=\; \int_\Theta \theta\,p(\theta|x)\,d\theta$$

the conditional expectation of the vector θ given x.

4.8.6 Example (Signal processing)**.** [39] STEVE & KEITH: This example(s) might have to go into the chapter on the Fourier transform.

[39] Further applications of variational methods to signal analysis can be found in sections 6.6 and ** FILL IN ** of *Signal Theory* by Lewis Franks, 1969, Prentice-Hall.

Let $H : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ be a bounded linear transformation given by the convolution

$$y(t) \;=\; Hx(t) \;=\; \int_{-\infty}^{\infty} h(t-s)\, x(s)\, ds \tag{4.8.1}$$

where $h \in L^2(\mathbb{R})$. H is a linear, time-invariant (LTI) filter in engineering talk, and H will be such a transformation (into L^2) if $h \in L^2(\mathbb{R})$. (Exercise: use Schwarz' inequality to prove this.) In the language of engineering the square of the L^2 norm of the 'signal' x is its 'power'

$$P(x) \;=\; \|x\|^2 \;=\; \int_{-\infty}^{\infty} |x(t)|^2\, dt\ ,$$

and an interesting question is: for a given LTI filter, what is the most power that can be obtained from that filter for any input signal x of unit power? That is, we are looking for the signal waveform (or shape) x which maximizes

$$\int_{-\infty}^{\infty} |y(t)|^2\, dt \;=\; \int_{-\infty}^{\infty} |Hx(t)|^2\, dt \tag{4.8.2}$$

subject to the constraint $\int_{-\infty}^{\infty} |x(t)|^2\, dt \;=\; 1$. (In a radar or sonar system maximizing the output power presumably allows the system to 'see' to its maximum range.)

Substituting (4.8.1) into (4.8.2), and using a Lagrange multiplier, means that we must maximize the quadratic functional

$$\begin{aligned} F(x) \;&=\; \int \Big| \int h(t-s)\, x(s)\, ds \Big|^2\, dt + \lambda(P(x)-1) \\ &=\; \int\int k(s_1,s_2)\ x(s_1)\, x(s_2)\, ds_1\, ds_2 + \lambda\Big(\int |x(t)|^2 dt - 1\Big) \end{aligned}$$

where the kernel

$$k(s_1,s_2) \;=\; \int h(t-s_1)\, h(t-s_2)\, dt$$

and all integrals are over the entire real line. The variation of F in the direction of u is

$$\delta F(x;u) \;=\; 2\int \Big[\int k(s,t)x(s)\, ds + \lambda\, x(t)\Big]\, u(t)\, dt\ .$$

Thus x satisfies

$$\int k(s,t)x(s)\, ds \;=\; -\lambda\, x(t)\ ,$$

that is, x is an eigenfunction for the integral operator with kernel k and the corresponding eigenvalue is $-\lambda$. If we multiply this equation by $x(t)$ and integrate over $\mathbb{R}$, and then use the constraint, we have

$$-\lambda \;=\; \int\int k(s,t)\, x(s)\, x(t)\, ds\, dt \;=\; \int |y(t)|^2\, dt\ .$$

This means that the value of the maximum power output of the system equals the largest eigenvalue for the kernel k, and the waveform that achieves this power is the corresponding eigenfunction. Integral operators, and their eigenstructure, will be treated in chapter **??**.

4.8.7 Example (Signal processing)**.** As in the preceding example we consider a linear, time-invariant filter $Hx(t) = \int_{-\infty}^{\infty} h(t-s)\, x(s)\, ds$. If $y(t) = Hx(t)$ is the output signal we desire to maximize the value $y(t_0)$ of the output at some time t_0 under the constraint that the power of the input is 1. This may be of interest if we wish to construct an output that has the best detection properties (maximum amplitude) at t_0. Since $y(t_0) = \int h(t_0 - s)\, x(s)\, ds$ we introduce a Lagrange multiplier and maximize

$$F(x) \;=\; \int h(t_0-s)\, x(s)\, ds + \lambda\Big(\int |x(s)|^2\, ds - 1\Big)\ .$$

where all integrals are over the entire real line. The variation in the direction u is

$$\delta F(x;u) \;=\; \int h(t_0-s)\,u(s)\,ds + 2\lambda \int x(s)\,u(s)\,ds$$

which shows that the necessary condition for a maximum is $x(t) = -h(t_0-t)/(2\lambda)$.

We can solve for λ by implementing the constraint $1 = \int x(t)^2\,dt = \int h(t_0-t)^2\,dt/(4\lambda^2)$. We obtain (taking the negative square root of λ^2)

$$x(t) \;=\; h(t_0-t)/\|h\|_{L^2}\ .$$

4.8.8 Exercise (Signal processing)**.** ** STOPPED HERE ** generalize the last example by using a more general linear functional (a detector)...

Can we derive the HFM sweep?

4.8.9 Example (Column buckling)**.** See C&H p 272+.

4.8.10 Example (Economic growth)**.** See Wan or Intriligator or Smith (p 16 and 48+).

4.8.11 Example (Geodesics on a manifold)**.**

4.8.12 Example (Rayleigh-Ritz procedure for eigenfunctions)**.**

4.8.13 Example (Route analysis)**.** In planning airline flight routes, cargo shipping routes, military troop movements, and other applications it is desired to find the most economical route for moving cargo from point A to point B. This may be similar to finding the shortest path, but there may be other factors to consider such as jet streams, ocean currents and weather, rough terrain, etc.

We find the path $x(t)$ from a to b in $\mathbb{R}^n$ that minimizes the total cost of moving from a to b where $x(0) = a$ and $x(1) = b$. The cost is measured by a functional

$$C(x) \;=\; \int_0^1 g(x(t))\,|\dot{x}(t)| + h(x(t))\,|\ddot{x}(t)|\,dt$$

where the first term in the integrand is an arclength or route cost, and the second is a speed of transit cost (e.g., burning fuel, etc.). The functions g and h are position dependent coefficients of difficulty or **HELP** such as terrain roughness, viscosity, etc. These coefficients are the real heart of the model!

Notice that if $h = 0$ and $g = 1$ the cost would be the arclength. *** STOPPED HERE ***

5 Optimal Estimation and Approximation in Hilbert Spaces

This section is directed toward problems that can be solved using projections onto finite dimensional subspaces of a Hilbert space. Many useful problems occur in Hilbert spaces of both finite and infinite dimension.[40]

5.1 Orthonormal Sets and Fourier Series

This section follows the exposition in Rudin, *Real and Complex Analysis*, chapter 4. Only separable Hilbert spaces will be considered; these are more intuitive and cover virtually all applications.

A metric space is *separable* if it has a countable dense subset. All finite dimensional vector spaces (over $\mathbb{R}$ or $\mathbb{C}$) are separable; the set of vectors with rational components is both countable and dense. The Hilbert spaces $\ell^2(\mathbb{N})$, $L^2(\Omega)$, and $H^1(\Omega)$ are also separable. The Banach space $L^\infty(\Omega)$ is not separable.

5.1.1 Exercise. Prove that a countable union of countable sets is countable. Use this to show that $\ell^2(\mathbb{N})$ is separable.

For notational convenience the set $\mathbb{N}$ will be the default index set in these notes. Other useful examples of countable index sets in applications are $\{1, ..., n\}$, $\mathbb{N}_0$, $\mathbb{Z}$, and $\mathbb{N}^k$ and $\mathbb{Z}^k$ for some $k \in \mathbb{N}$.

5.1.2 Definition. Let H be a (real or complex) Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and let $E = \{e_i \ ; \ i \in \mathbb{N}\}$ be a subset of H such that $e_i \neq 0$ for all $i \in \mathbb{N}$.

We say that E is *linearly independent* if every finite subset $\{e_1, ..., e_n\}$ of E is linearly independent as a set of vectors in H, i.e., if $\alpha_1 x_1 + \cdots + \alpha_n x_n = 0$ in H implies $\alpha_1 = \cdots = \alpha_n = 0$ in $\mathbb{K}$.

We say that E *spans* H if the set of all finite linear combinations of vectors in E is dense in H. (A finite linear combination of vectors in E is any vector of the form $\sum_{i=1}^n \alpha_i e_i$ where $n \in \mathbb{N}$, $e_i \in E$, and $\alpha_i \in \mathbb{K}$.)

We say that E is *orthonormal* if

$$\langle e_i, e_j \rangle \ = \ \begin{cases} 1 & \text{when } i = j \text{ in } \mathbb{N} \\ 0 & \text{when } i \neq j \text{ in } \mathbb{N} \end{cases} .$$

An orthonormal subset E is said to be *maximal* if there is no vector x in H such that the set $E \cup \{x\}$ is still orthonormal in H.

If E is an orthonormal subset and x is any element of H, the numbers $\langle e_i, x \rangle$, for $i \in \mathbb{N}$, are called the *Fourier coefficients* of x with respect to the set $\{e_i\}$. The Fourier coefficients of x are sometimes denoted by $\hat{x}_i$ if the orthonormal set is understood.

If the set of orthonormal vectors $\{e_i\}$, $i \in \mathbb{N}$, span H, then H is separable. For the set (when $\mathbb{K} = \mathbb{C}$)

$$\left\{ \sum_{i=1}^n \alpha_i e_i \ ; \ n \in \mathbb{N} \text{ and } \alpha_i \in \mathbb{Q} + i\mathbb{Q} \text{ for } i = 1, ..., n \right\}$$

is both countable and dense in H

5.1.3 Exercise. Prove this.

5.1.4 Theorem. *Let $e_1, e_2, ..., e_n$ be an orthonormal set in the Hilbert space H, let $c_k \in \mathbb{K}$ be scalars and set $x = \sum_1^n c_k \, e_k$. Then*

(i) $c_k \ = \ \langle e_k, x \rangle$,

(ii) $\|x\|^2 \ = \ \sum_1^n |c_k|^2$, *and*

(iii) the set $\{e_k\}_1^n$ is linearly independent.

[40] TO STEVE: I should include a section on 'non-linear' least squares; this is useful and there is not a whole lot that needs be said, just algorithms.

Proof. With x so given we have[41] $\langle e_k, x\rangle = \sum_{\ell=1}^n c_\ell \langle e_k, e_\ell\rangle = \sum_{\ell=1}^n c_\ell \delta_{k\ell} = c_k$. And we may compute $\|x\|^2 = \langle x, x\rangle = \sum_{k=1}^n \sum_{\ell=1}^n \bar{c}_k c_\ell \langle e_k, e_\ell\rangle = \sum_{k=1}^n \sum_{\ell=1}^n \bar{c}_k c_\ell \delta_{k\ell} = \sum_{k=1}^n \bar{c}_k c_k$. The linear independence of $\{e_k\}$ follows from (ii). □

5.1.5 Corollary. *Let $\{e_i\ ;\ i \in \mathbb{N}\}$ be an orthonormal set. Then this set is linearly independent.*

5.1.6 Theorem. *Let $e_1, e_2, ..., e_n$ be an orthonormal set in the Hilbert space H and let $x \in H$ be any vector. Then*

(i) $\|x - \sum_{k=1}^n \langle e_k, x\rangle\, e_k\| \ \leq\ \|x - \sum_{k=1}^n c_k\, e_k\|$ *for all $c_1, ..., c_n$ in $\mathbb{K}$, and equality holds if and only if $c_k = \langle e_k, x\rangle$ for all $k = 1, ..., n$.*

(ii) $\sum_{k=1}^n \langle e_k, x\rangle\, e_k$ *is the orthogonal projection of x onto the span of $\{e_1, ..., e_n\}$.*

(iii) If $\delta = \text{dist}(x, \text{span}\{e_k\})$ *then* $\|x\|^2 = \sum_1^n |\langle e_k, x\rangle|^2 + \delta^2$.

Proof. Let M be the span of $\{e_1, ..., e_n\}$ in H. By Corollary **??** M is closed. According to Theorem **??**, if we let Px denote the projection of x onto M and set $y = Px$ we have $\|x - y\| \leq \|x - z\|$ for all $z \in M$. Furthermore, $Qx = y - x \in M^\perp$ so $\langle x - y, z\rangle = 0$ for all $z \in M$. In particular,

$$\langle x - y, e_k\rangle \ =\ 0 \tag{5.1.1}$$

for all $k = 1, ..., n$.

Now let's set $y = \sum_1^n c_k\, e_k$ for some unknown coefficients c_k. Then (5.1.1) implies that

$$\langle e_k, x\rangle \ =\ \langle e_k, \sum_{\ell=1}^n c_\ell\, e_\ell\rangle \ =\ \sum_{\ell 1}^n c_\ell \delta_{k\ell} \ =\ c_k\ .$$

This shows (i) and (ii).

Finally, let δ be $\|x - y\|$. Then

$$\delta^2 \ =\ \langle x - y, x - y\rangle \ =\ \langle x - y, x\rangle \ =\ \langle x, x\rangle - \langle \sum_{k=1}^n \langle e_k, x\rangle e_k, x\ \rangle \ =\ \|x\|^2 - \sum_{k=1}^n \overline{\langle e_k, x\rangle}\, \langle e_k, x\rangle\ .$$

□

We now come to two important relationships, Bessel's inequality and Parseval's equality.

5.1.7 Corollary (Bessel's inequality)**.** *Let $\{e_i\ ;\ i \in \mathbb{N}\}$ be an orthonormal set in the Hilbert space H. And let $x \in H$ and $\hat{x}_i \in \mathbb{K}$ be the Fourier coefficients of x. Then*

$$\sum_{i=1}^\infty |\hat{x}_i|^2 \ \leq\ \|x\|^2\ .$$

5.1.8 Theorem. *Let $\{e_i\ ;\ i \in \mathbb{N}\}$ be an orthonormal set in the Hilbert space H. Then the following properties are equivalent.*

(i) $\{e_i\ ;\ i \in \mathbb{N}\}$ *is maximal in H;*

(ii) $\{e_i\ ;\ i \in \mathbb{N}\}$ *spans H;*

(iii) $\sum_{i\in\mathbb{N}} |\langle e_i, x\rangle|^2 \ =\ \|x\|^2$ *for all $x \in H$ (Parseval's equality);*

(iv) $\langle x, y\rangle \ =\ \sum_{i\in\mathbb{N}} \overline{\langle e_i, x\rangle}\, \langle e_i, x\rangle$ *for all $x, y \in H$ (Parseval's identity).*

[41] TO STEVE: I conjugate the first entry of the inner product, here.

An orthonormal set $\{e_i\}$ in a Hilbert space H is said to be *complete*, or an *orthonormal basis*, if it has any one, and hence all, of the properties listed in Theorem 5.1.8. We may write c.o.n.s for complete orthonormal set.

Proof. (i)⇒(ii). Denote by E the subspace of H consisting of all finite linear combinations of elements of $\{e_i\ ;\ i \in \mathbb{N}\}$. The closure, $\bar{E}$, of E is also a subspace of H (we leave to the reader the verification that both E and $\bar{E}$ are subspaces). Since $\{e_i\ ;\ i \in \mathbb{N}\}$ is maximal there is no vector $x \in H$ which is orthogonal to every e_i. But if $\bar{E} \neq H$ there is a nonzero $x \in \bar{E}^{\perp}$ contradicting the maximality of $\{e_i\ ;\ i \in \mathbb{N}\}$.

(ii)⇒(iii). Letting E denote the same subspace as above, we assume E is dense in H. Pick $x \in H$ and $\epsilon > 0$. Then we can find a finite set $i_1, ..., i_n$ in $\mathbb{N}$, and scalars $\alpha_1, ..., \alpha_n$, such that $\|x - \sum_1^n \alpha_k\, e_{i_k}\| < \epsilon$. By Theorem 5.1.6 this bound is only improved if we replace each α_k by $\langle e_{i_k}, x\rangle$, so that $\|x - \sum_1^n \langle e_{i_k}, x\rangle\, e_{i_k}\| < \epsilon$. This implies that

$$\|x\| \ \leq\ \|\sum_1^n \langle e_{i_k}, x\rangle\, e_{i_k}\| \ + \|x - \sum_1^n \langle e_{i_k}, x\rangle\, e_{i_k}\| \ <\ \|\sum_1^n \langle e_{i_k}, x\rangle\, e_{i_k}\| + \epsilon\ .$$

And this inequality can be changed to give

$$(\|x\| - \epsilon)^2 \ <\ \|\sum_1^n \langle e_{i_k}, x\rangle\, e_{i_k}\|^2 \ =\ \sum_1^n |\langle e_{i_k}, x\rangle|^2\ .$$

By Bessel's inequality this last expression is bounded by $\|x\|^2$; since ϵ is arbitrarily small (5.1) shows (iii).

(iii)⇒(iv). Let x and y be elements of H with Fourier coefficients $\hat{x}_i$ and $\hat{y}_i$ (and complex conjugates $\bar{\hat{x}}_i$ and $\bar{\hat{y}}_i$), and let $\alpha \in \mathbb{K}$. Then $x + \alpha y$ hss Fourier coefficients $\hat{x}_i + \alpha\hat{y}_i$, and by assumption

$$\langle x + \alpha y, x + \alpha y\rangle \ =\ \sum_{i\in\mathbb{N}} (\overline{\hat{x}_i + \alpha\hat{y}_i})(\hat{x}_i + \alpha\hat{y}_i)\ .$$

Expanding the inner product on the left and multiplying out the terms on the right, and then using (iii) to cancel the terms $\|x\|$, $\|y\|$, and their Fourier coefficients, gives

$$\alpha\langle x, y\rangle + \bar{\alpha}\langle y, x\rangle \ =\ \alpha \sum_{i\in\mathbb{N}} \bar{\hat{x}}_i\hat{y}_i + \bar{\alpha}\sum_{i\in\mathbb{N}} \hat{x}_i\bar{\hat{y}}_i\ .$$

Setting $\alpha = 1$ shows that $\Re\langle x, y\rangle = \Re\sum_{i\in\mathbb{N}} \bar{\hat{x}}_i\hat{y}_i$, and (if $\mathbb{K} = \mathbb{C}$) setting $\alpha = i$ shows that $\Im\langle x, y\rangle = \Im\sum_{i\in\mathbb{N}} \bar{\hat{x}}_i\hat{y}_i$.

(iv)⇒(i). Of course (vi) trivially implies (iii) by setting $y = x$. Now assume (i) is false, so that there exists an $x \in H$, $x \neq 0$, such that $\langle e_i, x\rangle = 0$ for all $i \in \mathbb{N}$. But then (iii) immediately gives $\|x\| = 0$ which is a contradiction. □

5.1.9 Theorem. *If H is any separable Hilbert space, there is at least one countable, complete orthonormal set in H. If H is finite dimensional, of dimension n, any such set will, of course, have n vectors.*

Proof. If H is separable there is a countable dense subset $x_1, x_2, ...$ By applying the Gram-Schmidt process to this subset, and deleting vectors which are not linearly independent, we arrive at a complete orthonormal set. The details have already been given in the proof of Theorem 1.6.8. □

The following result is sometimes useful, for instance in studying partial differential equations with both time and space variables, or stochastic processes which have both time and sample space variables.

5.1.10 Theorem. *Let Ω_1 and Ω_2 be open subsets of $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ respectively, and let u_i, $i \in \mathbb{N}$, and v_j, $j \in \mathbb{N}$, be complete orthonormal sets for $L^2(\Omega_1)$ and $L^2(\Omega_2)$ respectively. Then the functions $w_{ij}(x, y) = u_i(x)\, v_j(y)$, $i, j \in \mathbb{N}$, form a complete orthonormal set for $L^2(\Omega_1 \times \Omega_2)$.*

Proof. We leave this proof as an exercise for students with some knowledge of the Lebesgue spaces $L^2(\Omega)$. □

5.1.11 Exercise. Let $H = \mathbb{R}^m$ and $K = \mathbb{R}^n$. Let $u_1, ..., u_m$ be an orthonormal basis for H and $v_1, ...v_n$ an orthonormal basis for K. Then $H \otimes K$ is (isomorphic to) the vector space of all $m \times n$ matrices (with real components), and a basis for this (mn dimensional) vector space is the set of all $u_i \otimes v_j$, $1 \le i \le m$ and $1 \le j \le n$. (If u_i and v_j denote column vectors, $u_i \otimes v_j = u_i\, v_j^t$, ^{t}=transpose.)

5.1.12 Exercise. Let $\Omega \subset \mathbb{R}^d$ be open and $v_i(x)$ be orthonormal basis for $L^2(\Omega)$, and let e_j, $j = 1, ..., n$, be an orthonormal basis for $\mathbb{K}^n$ where $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. Show that the set of vector valued function $u_{ij}(x) = v_i(x)\, e_j$ is an orthonormal basis for $L^2(\Omega; \mathbb{K}^n)$.

Completeness of the trigonometric functions in L^2 We show that the classical trigonometric system $\{e^{2\pi ikx}\ ;\ k \in \mathbb{Z}\}$ is complete in $L^2((0,1); \mathbb{C})$. Similar theorems can be shown for other L^2 spaces, when $(0,1)$ is replaced by (a,b), or for other systems of functions, such as $\{\cos(kx), \sin(kx)\ ;\ k \in \mathbb{N}_0\}$. Proofs may be constructed by imitating the proof of the following theorem, or by using this result with a change of the x variable.

5.1.13 Theorem. *The following sets of functions are orthonormal and complete in the indicated Hilbert space (in general over $\mathbb{C}$).*

- $\{e^{2\pi inx}\ ;\ n \in \mathbb{Z}\}$ *in* $L^2(0,1)$
- $\{\frac{e^{inx}}{\sqrt{2\pi}}\ ;\ n \in \mathbb{Z}\}$ *in* $L^2(0, 2\pi)$
- $\{\frac{1}{\sqrt{2\pi}}, \frac{\sin nx}{\sqrt{2\pi}}, \frac{\cos nx}{\sqrt{2\pi}}\ ;\ n \in \mathbb{N}\}$ *in* $L^2(0, 2\pi)$
- $\{e^{2\pi i(mx+ny)}\ ;\ m, n \in \mathbb{Z}\}$ *in* $L^2((0,1)^2)$
- $\{\frac{e^{i(mx+ny)}}{2\pi}\ ;\ m, n \in \mathbb{Z}\}$ *in* $L^2((0, 2\pi)^2)$
- $\{\frac{e^{i(\mathbf{n}\cdot\mathbf{x})}}{(2\pi)^{d/2}}\ ;\ \mathbf{n} \in \mathbb{Z}^d\}$ *in* $L^2((0, 2\pi)^d)$, *where* $\mathbf{n} \cdot \mathbf{x} = n_1 x_1 + \cdots + n_d x_d$

The intervals $(0,1)$ and $(0, 2\pi)$ may be replaced by any other intervals of length 1 and 2π, respectively.

The proof is substantially contained in section 2.2 of Dettman. A nice proof can also be found in section 10.3 (proofs at end of chapter) of Marsden *ECA*.

5.1.14 Exercise. Prove that the functions $\{e^{2\pi inx}\ ;\ n \in \mathbb{N}\}$ are orthonormal in $L^2((0,1); \mathbb{C})$.

5.1.15 Exercise. Can the functions $\{e^{2\pi inx}\ ;\ n \in \mathbb{N}\}$ be turned into an orthonormal set in $H^1((0,1); \mathbb{C})$? Show how.

5.1.16 Exercise. Use Gram-Schmidt to orthonormalize the set $\{1, x, x^2, x^3, ..., x^n\}$ in $L^2(a,b)$. Calculate the first three terms explicitly.

5.1.17 Exercise. Use Gram-Schmidt to orthonormalize the set $\{1, x, x^2, x^3, ..., x^n\}$ in $H^1(a,b)$. Calculate the first three terms explicitly.

5.1.18 Exercise. Use Gram-Schmidt to orthonormalize the set $\{1, x, x^2, x^3, ..., x^n\}$ in $L^2(\mathbb{R}; e^{-x^2/2}dx)$. Calculate the first three terms explicitly. These are the Hermite polynomials.

5.1.19 Exercise. Use Gram-Schmidt to orthonormalize the set $\{1, x, y, x^2, xy, y^2\}$ in $L^2((a,b) \times (c,d))$.

5.1.20 Exercise. Let $n \in \mathbb{N}$ and, for $i = 1, ..., n$, $I_i = (\frac{i-1}{n}, \frac{i}{n})$ be a partition of the interval $(0,1)$. Use Gram-Schmidt to orthonormalize the set of functions $1_{I_i}(x)$ in $L^2(0,1)$.

5.2 Least Squares Models

In Theorem **??** we established an explicit formula for the projection P onto any finite dimensional subspace of a Hilbert space. In $\mathbb{C}^n$ this formula is $P = X(X^*X)^{-1}X^*$ where the columns of the $n \times m$ matrix X form a basis for the m-dimensional subspace of $\mathbb{C}^n$. The same formula holds in $\mathbb{R}^n$ if X^* is replaced by X'.

In an infinite dimensional Hilbert space the same formula still holds provided that we interpret the ij-th entry of the $m \times m$ Gram matrix X^*X (or $X'X$) as the Hilbert space inner product of the i-th and j-th vectors which make up the 'columns' of X. To be explicit, if $y \in H$ and we want to find the projection $\hat{y}$ of y onto the span of $x_1, ..., x_m$, the formula is

$$\hat{y} = \sum_1^m \beta_j x_j \qquad \text{where the } \beta\text{'s satisfy} \qquad \begin{pmatrix} \langle x_1, x_1\rangle & \cdots & \langle x_1, x_m\rangle \\ \vdots & & \vdots \\ \langle x_m, x_1\rangle & \cdots & \langle x_m, x_m\rangle \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} = \begin{pmatrix} \langle x_1, y\rangle \\ \vdots \\ \langle x_m, y\rangle \end{pmatrix}.$$

Note that if the x_j's are orthogonal, the Gram matrix is diagonal and we are left with

$$\hat{y} \;=\; \sum_1^m \left(\frac{\langle x_j, y\rangle}{\|x_j\|^2}\right) x_j\,,$$

the m-th partial sum of the Fourier series for y in the orthonormal set $x_j/\|x_j\|$.[42]

We now continue our discussion of projections by giving some applications.

Linear regression

5.2.1 Example (fitting a line to real data)**.** We determine the equation of a line that best fits n points $(x_j, y_j), j = 1, ..., n$, in $\mathbb{R}^2$. We choose the slope b and y-intercept a of the function $f(x) = ax + b$ so that the sum of squares

$$\sum_{j=1}^n |y_j - f(x_j)|^2 \;=\; \sum_{j=1}^n |y_j - (a + bx_j)|^2 \tag{5.2.1}$$

is minimized. It is quite possible to differentiate this expression with respect to a and b and derive the following formulas, but we will use the normal equations. The model equations are

$$y_j = a + bx_j + e_j$$

for $j = 1, ..., n$ where e_j is an error. In matrix form we have

$$y \;=\; X\beta + e$$

where $\beta = (a\ b)'$, $y = (y_1, ..., y_n)'$, $e = (e_1, ..., e_n)'$, and the design matrix

$$X \;=\; \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{n\times 2}.$$

The normal equations are

$$X'X\hat{\beta} \;=\; \begin{pmatrix} n & \sum_1^n x_j \\ \sum_1^n x_j & \sum_1^n x_j^2 \end{pmatrix} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} \;=\; \begin{pmatrix} \sum_1^n y_j \\ \sum_1^n x_j y_j \end{pmatrix} \;=\; X'y$$

where the notation $\hat{a}$, $\hat{b}$ is used for the minimizing a and b. Explicitly we have

$$\begin{aligned} \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} \;&=\; \frac{1}{n\sum_1^n x_j^2 - (\sum_1^n x_j)^2} \begin{pmatrix} \sum_1^n x_j^2 & -\sum_1^n x_j \\ -\sum_1^n x_j & n \end{pmatrix} \begin{pmatrix} \sum_1^n y_j \\ \sum_1^n x_j y_j \end{pmatrix} \\ &=\; \frac{1}{n\sum_1^n x_j^2 - (\sum_1^n x_j)^2} \begin{pmatrix} (\sum_1^n x_j^2)(\sum_1^n y_j) - (\sum_1^n x_j)(\sum_1^n x_j y_j) \\ n(\sum_1^n x_j y_j) - (\sum_1^n x_j)(\sum_1^n y_j) \end{pmatrix} \end{aligned}$$

where every sum is over j.

[42] Note that if these x_j's are column vectors in $\mathbb{C}^n$ this a weighted sum of dyads $\hat{y} = \sum_1^m \frac{1}{\|x_j\|^2} x_j x_j^* y$.

5.2.2 Example (fitting a polynomial to real data)**.** Assume we are given n pairs of real numbers (x_j, y_j) and we wish to find the polynomial of degree $(k-1)$ which is closest to our data in the least squares sense. We assume $k \leq n$. A model may be written

$$y_j \;=\; \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \cdots + \beta_{k-1} x_j^{k-1} + e_j$$

where e_j is the model error. We put into the design matrix one column for every term in this model, that is, one column for every unknown coefficient we want to fit. Thus, set

$$X \;=\; \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{k-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{k-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^{k-1} \end{pmatrix}_{n\times k} .$$

With all of the y-data placed in the column vector $y = (y_1, ..., y_n)'$, and all errors assembled into the $n \times 1$ column vector e, we can write the model in matrix form

$$y \;=\; X\beta + e \tag{5.2.2}$$

where $\beta = (\beta_0, ..., \beta_{k-1})'$. If at least k of the x_j's are distinct, X will be full rank (but we will not prove this) and the minimizing $\hat{\beta}_j$'s can be obtained from the normal equations

$$(X'X)\hat{\beta} \;=\; X'y \qquad \text{or} \qquad \hat{\beta} \;=\; (X'X)^{-1}X'y \; . \tag{5.2.3}$$

5.2.3 Example (fitting a plane to real data)**.** Assume we are given n, $(k+1)$-tuples of real numbers

$$(x_{1j}, x_{2j}, ..., x_{kj}, y_j) \;\in\; \mathbb{R}^{k+1} \qquad\qquad j = 1, ..., n$$

and we wish to find the affine hyper-plane in the independent variables $x_1, ..., x_k$ which is closest to our data in the least squares sense. The model is

$$y_j \;=\; \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj} + e_j$$

and we want to minimize the sum of squares of the errors

$$\sum_{j=1}^{n} e_j^2 \;=\; \sum_{j=1}^{n} (y_j - \beta_0 - \beta_1 x_{1j} - \beta_2 x_{2j} - \cdots - \beta_k x_{kj})^2 \; .$$

We require $k+1 \leq n$, and put into the design matrix one column for every unknown coefficient we want to fit:

$$X \;=\; \begin{pmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \dots & x_{kn} \end{pmatrix}_{n\times(k+1)} .$$

Putting all of the y-data in a column vector $y = (y_1, ..., y_n)'$ and the errors in a column vector e the model equations take the form (5.2.2) where X is $n \times (k+1)$ and β is $(k+1) \times 1$. If there are k linearly independent vectors $(x_{1j}, x_{2j}, ..., x_{kj})$ in this set of n points in $\mathbb{R}^k$, the matrix X will be full rank. The best model parameters $\beta = (\beta_0, \beta_1, ..., \beta_k)'$ are again obtained from the normal equations (5.2.3).

5.2.4 Example (fitting a trigonometric sum)**.** Assume we are given n pairs of real numbers (x_j, y_j) and we wish to find the r-th degree trigonometric polynomial which is closest to our data in the least squares sense. A model may be written

$$y_j \;=\; \beta_0 + \sum_{k=1}^{r} \Big(\beta_{2k-1} \sin(kx_j) + \beta_{2k} \cos(kx_j) \Big) + e_j$$

where e_j is the error. We require $2r+1 \le n$. A design matrix is

$$X \;=\; \begin{pmatrix} 1 & \sin(x_1) & \cos(x_1) & \dots & \sin(rx_1) & \cos(rx_1) \\ 1 & \sin(x_2) & \cos(x_2) & \dots & \sin(rx_2) & \cos(rx_2) \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & \sin(x_n) & \cos(x_n) & \dots & \sin(rx_n) & \cos(rx_n) \end{pmatrix}_{n\times(2r+1)} .$$

With the data y_j and errors e_j in their respective $n \times 1$ vectors y and e we can write this system in the form (5.2.2) and find the best model parameters $\beta = (\beta_0, ..., \beta_{2r+1})'$ from the normal equations (5.2.3), assuming X is full rank.

5.2.5 Exercise. Using the same data and model as Example 5.2.1, set $x = (x_1, ..., x_n)'$, $y = (y_1, ..., y_n)'$, $\overline{x} = (\sum_1^n x_j)/n$, and $\overline{y} = (\sum_1^n y_j)/n$. Show that

$$(X'X)^{-1} \;=\; \frac{1}{|x|^2 - n\,\overline{x}^2} \begin{pmatrix} |x|^2/n & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix} , \qquad \text{and} \qquad X'y \;=\; \begin{pmatrix} n\,\overline{y} \\ x \cdot y \end{pmatrix} ,$$

and give an expression for $\hat{a}$ and $\hat{b}$ using these variables.

5.2.6 Exercise. Work out the analogous formulas in Example 5.2.1 for $\hat{a}, \hat{b} \in \mathbb{C}$ when the x_j's and y_j's are complex numbers. Show that the complex model is equivalent to a real model with twice the data and twice as many parameters.

5.2.7 Exercise. Suppose the x and y axes in $\mathbb{R}^2$ are shifted so that $\overline{x} = \overline{y} = 0$ in Example 5.2.1. Give the formula for the least squares parameters $\hat{\beta} = (\hat{a}, \hat{b})'$ in this case. How does this simplification constrain the data points (x_i, y_i) in $\mathbb{R}^2$? Show that every least squares problem of the form (5.2.1) can be reduced to this case by such a shift. How are the true unknown parameters a and b changed by such a shift?

5.2.8 Exercise. Why can we not allow $k > n$ in Example 5.2.2? Show that if $k = n$ the least squares polynomial (of degree $n-1$) will interpolate the n pairs of data points.

5.2.9 Exercise. Compute the ij-th component of the Gram matrix $X'X$ for Examples 5.2.2 and 5.2.3. Also compute the components of the vector $X'y$.

5.2.10 Exercise. Do this exercise by hand. Fit a least squares line $y = \beta_0 + \beta_1 x$ to the three data points $(-1, 0), (0, 2), (1, 2)$ in the xy-plane. In matrix notation we have $\mathbf{y} = X\beta + \mathbf{e}$ where $\mathbf{y} = (0, 2, 2)'$, $\beta = (\beta_0, \beta_1)'$, and $X = [\mathbf{x}_1, \mathbf{x}_2]$ with $\mathbf{x}_1 = (1, 1, 1)'$ and $\mathbf{x}_2 = (-1, 0, 1)'$. Draw a three dimensional picture of $\mathbf{y}$, $\mathbf{x}_1$, and $\mathbf{x}_2$, and then plot the projection of $\mathbf{y}$ on the span of $\mathbf{x}_1$ and $\mathbf{x}_2$.

On a two dimensional graph, plot the three (x_i, y_i) data points. Compute $X'X$, $(X'X)^{-1}$, $X'\mathbf{y}$, and the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Compute the fitted values $\hat{\mathbf{y}} = X\hat{\beta}$, and plot the three pairs of points $(x_i, \hat{y}_i)$ on the graph. Also plot the regression line $y = \hat{\beta}_0 + \hat{\beta}_1 x$ on the graph. Identify the fitted errors $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$ on both the three dimensional picture and the two dimensional graph.

5.2.11 Exercise. Do this exercise on a computer. Fit a quadratic $y = \beta_0 + \beta_1 x + \beta_2 x^2$ to the data $(-2, -1), (-1, 0.5), (-1, 1), (0, 1), (1, 1.5), (2, 0.5), (2.1), (3, 1)$ using least squares. On an xy-plane, plot the data, the regression curve, and the fitted values $(x_i, \hat{y}_i)$ for $i = 1, ..., 8$. Identify the errors of this least squares fit.

5.2.12 Exercise (complex data)**.** Consider the linear model $y = X\beta + e$ when y and e are in $\mathbb{C}^n$, $\beta \in \mathbb{C}^m$ with $m \le n$, and X has only real entries. Show that the linear model is equivalent to two real models

$$\Re y \;=\; X\,\Re\beta + \Re e \qquad \text{and} \qquad \Im y \;=\; X\,\Im\beta + \Im e \; .$$

Suppose that all variables, y, X, β, and e, are complex valued with $y = u + iv$, $X = A + iB$, $\beta = \mu + i\nu$, and $e = \xi + i\eta$. Show that the linear model is equivalent to the $2n$ system of linear equations

$$\begin{pmatrix} u \\ v \end{pmatrix} \;=\; \begin{pmatrix} A & -B \\ B & A \end{pmatrix} \begin{pmatrix} \mu \\ \nu \end{pmatrix} + \begin{pmatrix} \xi \\ \eta \end{pmatrix} \; .$$

Show that X is full rank if and only if the real design matrix

$$\begin{pmatrix} A & -B \\ B & A \end{pmatrix}$$

is full rank.

5.2.13 Exercise. What conditions on the set x_j, $j = 1, ..., n$, in the previous example will ensure that X is full rank?

5.2.14 Exercise. Suppose we have n complex numbers $z_1, z_2, ..., z_n$ that were produced by a noisy system, the value z_j being produced at time t_j. And suppose we want to model this system by an equation of the form

$$z_j \;=\; \sum_{k=1}^{m} a_k \, e^{i\,\omega_k\, t_j} + e_j$$

where $\omega_1, ..., \omega_m$ are distinct known real frequencies, $a_1, ..., a_m$ are unknown complex amplitudes, and $e_1, ..., e_n$ are unknown complex errors which represent noise. Assuming $m \leq n$, give the design matrix for this model, and compute the jk-th entry of the Gram matrix X^*X and the k-th component of the vector X^*z. Here we use the matrix notation $z = Xa + e$.

5.2.15 Exercise. Write out the design matrix to fit the model

$$y_i \;=\; b_{00} + b_{10}x_{1i} + b_{01}x_{2i} + b_{20}x_{1i}^2 + b_{11}x_{1i}x_{2i} + b_{02}x_{2i}^2 + e_i$$

to a set of data $\{(x_{1i}, x_{2i}, y_i)\}_{i=1}^n$. Here the e_i are model errors, the x_1 and x_2 are independent variables, and the parameters b_{jk} are to be determined. Also give the Gram matrix $X'X$, and the vector $X'y$.

Analysis of variance and categorical models

5.2.16 Example (one-way layout)**.** Analysis of variance (ANOVA) models use categorical x's to predict the value of y. If there are p categories with n_i observations in category i, the one-way model for the j-th observation in the i-th category is

$$y_{ij} \;=\; \mu_i + e_{ij}$$

for $i = 1, ..., p$ and $j = 1, ..., n_i$. A typical (and very basic) null statistical hypothesis is $H : \mu_1 = \mu_2 = \cdots = \mu_p$ with the alternate hypothesis being that at least one of the categories has a population mean different than the others.

For instance, if we have observations on beer drinkers who drive sports cars, or fish, or like machine learning, the question might be whether the three groups drink about the same amount of beer, or whether one or two of the groups drink more than the other(s).

The model can be writen in the form $y = X\beta + e$ where

$$y \;=\; \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{p1} \\ \vdots \\ y_{pn_p} \end{pmatrix}, \qquad X \;=\; \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \qquad \text{and} \quad \beta \;=\; \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}.$$

The vector y is $n \times 1$ where $n = n_1 + \cdots + n_p$.

5.2.17 Example (two-way layout)**.** In these models there are two sets of categorical variables that help predict the value of y. For instance, in agricultural experiments on corm productivity one may be interested in brand of fertilizer used, A, B, C, D, as well as soil type I, II, III. Part of the concept is that the same brand of fertilizer may not be best for all soil types.

We might model y, the corn yield per acre, as

$$y_{ijk} \;=\; \mu_i + \nu_j + e_{ijk}\,, \qquad i = 1, ..., I \text{ and } j = 1, ..., J\,,$$

where we think of μ_i as the mean yield using the i-th fertilizer, averaged over all soil types, and ν_j as the mean of the j-th soil type, averaged over all brands of fertilizer. The index k identifies the k field that had soil type j and was given fertilizer i.

5.2.18 Exercise. In a study of 'human compasion' n subjects were given a questionaire, the over-all score of which is supposed to be a measure of their ability to feel compasion. One of the background questions on the questionaire was the degree of education the subject had. The education choices were: less than high school, high school, college, and graduate school or other post college degree. In each of these four categories there were n_1, n_2, n_3, and n_4 responses, with $n = n_1 + n_2 + n_3 + n_4$. Write out an analysis-of-variance model of this situation.

5.2.19 Exercise. Write out and solve the normal equations to give an explicit formula for each $\hat{\mu}_i$ in the one-way layout. How do your formulas compare with the sample average as an estimate of the mean for a single probability density? Is the sample mean of a random sample from a single population (probability density) a least squares estimate?

5.2.20 Exercise. Write out a design matrix, of 0's and 1's, for the two-way layout. Is your matrix full rank? An alternate set of equations for the two-way layout is

$$y_{ijk} \;=\; \mu + \alpha_i + \beta_j + e_{ijk}$$

where μ is now an over-all mean and the α's and β's are off-sets from this over-all mean, and are subject to constraints $\alpha_1 + \cdots + \alpha_I = 0$ and $\beta_1 + \cdots + \beta_J = 0$. How does the design matrix change for this set of parameters? Is it full rank?

Auto-regressive time series models

5.2.21 Example. Consider a second order auto-regressive model

$$y_k \;=\; a_1\, y_{k-1} + a_2\, y_{k-2} + e_k$$

of the time series y_k, $k = 1, ..., n$. The model coefficients a_1 and a_2 are to be estimated and e_k is a random error. If we have observed the real data $y_1, y_2, ..., y_n$ we can set up a regression model $y = X\beta + e$ as

$$\begin{pmatrix} y_n \\ y_{n-1} \\ \vdots \\ y_4 \\ y_3 \end{pmatrix} = \begin{pmatrix} y_{n-1} & y_{n-2} \\ y_{n-2} & y_{n-3} \\ \vdots & \vdots \\ y_3 & y_2 \\ y_2 & y_1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} e_n \\ e_{n-1} \\ \vdots \\ e_4 \\ e_3 \end{pmatrix}.$$

The estimate of a_1 and a_2 is made using the normal equations.

5.2.22 Exercise. Write out the matrix equation $y = X\beta + e$ for the auto-regressive model

$$y_k \;=\; a_1\, y_{k-1} + a_2\, y_{k-2} + \cdots + a_p\, y_{k-p} + e_k\,.$$

Write out the normal equations for this model.

5.2.23 Exercise. Consider the data $y_1, y_2, ..., y_n$ and the model in Example 5.2.21. Suppose now that a few of the observations y_k are missing from the set of n. Can a_1 and a_2 still be estimated? How does the design matrix X change?

Multivariate models The least squares models work just as well when the observations y_j take their values in $\mathbb{R}^m$ or $\mathbb{C}^m$. This extension is almost trivial but very useful in certain applications, particularly when the same independent variables (sometimes called *explanatory variables*) $x_1, ..., x_k$ can be used to predict several dependent variables (sometimes called *response variables*) $y_1, ..., y_m$. We now consider m models of the type previously considered

$$\begin{aligned} y_1 &= x_1 b_{11} + \cdots + x_k b_{k1} + e_1 \qquad \text{(model 1)} \\ y_2 &= x_1 b_{12} + \cdots + x_k b_{k2} + e_2 \qquad \text{(model 2)} \\ &\vdots = \vdots \\ y_m &= x_1 b_{1m} + \cdots + x_k b_{km} + e_m \qquad \text{(model } m\text{)} \ . \end{aligned}$$

Each model is distinguished by its own set of parameters, $b_{1j}, ..., b_{kj}$ for the j-th model. As with the single model the problem of model building is to estimate values for these model parameters using the data of some experiment. *The least squares minimization problem here is really m separate minimization problems.* For each model we assume we have n observations, each of the form $(x_1, ..., x_k, y)$. Thus, our data set looks like

$$\begin{aligned} &(x_{11}, x_{12}, ..., x_{1k}, y_{1j}) \\ &(x_{21}, x_{22}, ..., x_{2k}, y_{2j}) \\ &\qquad \cdots \\ &(x_{n1}, x_{n2}, ..., x_{nk}, y_{nj}) \end{aligned}$$

for $j = 1, ..., m$. The explanatory variables $x_{i1}, ..., x_{ik}$ do not depend on j, the model index. For each set of k variables $x_{i1}, ..., x_{ik}$, i.e., for each $i = 1, ..., n$, we have m dependent, or response, variables $y_{i1}, ..., y_{im}$.

We write all data and parameters as a system of linear equations

$$y_{ij} = x_{i1} b_{1j} + \cdots + x_{ik} b_{kj} + e_{ij} \qquad i = 1, ..., n \quad \text{and} \quad j = 1, ..., m \tag{5.2.4}$$

where e_{ij} is the model mismatch or error for the data value y_{ij}. These equations can be put into matrix form

$$Y = XB + E \tag{5.2.5}$$

where

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nm} \end{pmatrix}, \qquad X = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix},$$

$$B = \begin{pmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & & \vdots \\ b_{k1} & \cdots & b_{km} \end{pmatrix}, \qquad E = \begin{pmatrix} e_{11} & \cdots & e_{1m} \\ \vdots & & \vdots \\ e_{n1} & \cdots & e_{nm} \end{pmatrix}.$$

This is a compact way of writing all the data for the m models when each model uses the same design matrix X.

The normal equations for the models, i.e., for B, can be written simultaneously as

$$(X'X)B = X'Y \qquad \text{or} \qquad B = (X'X)^{-1}X'Y \ . \tag{5.2.6}$$

5.2.24 Example. Suppose it is required to predict certain medical conditions in patients based on patient data. The prediction is useful in situations where the medical conditions are subtle or difficult to diagnose and the data is, in contrast, easy to obtain.

Consider the response variables

$$\begin{aligned} y_1 &= \text{risk for heart disease} \\ y_2 &= \text{risk for skin cancer} \\ y_3 &= \text{risk for prostate cancer} \\ y_4 &= \text{risk for breast cancer} \\ y_5 &= \text{risk for diabetes} \\ y_6 &= \text{hearing loss} \\ y_7 &= \text{eye sight ,} \end{aligned}$$

and the explanatory variables

$$\begin{aligned} x_1 &= \text{age} \\ x_2 &= \text{height} \\ x_3 &= \text{weight} \\ x_4 &= \text{gender} \quad \text{(female=1, male=0)} \\ x_5 &= \text{father's history of heart disease (yes=1 or no=0)} \\ x_6 &= \text{father's history of skin cancer (yes=1 or no=0)} \\ x_7 &= \text{father's history of prostate cancer (yes=1 or no=0)} \\ x_8 &= \text{father's history of diabetes (yes=1 or no=0)} \\ x_9 &= \text{mother's history of heart disease (yes=1 or no=0)} \\ x_{10} &= \text{mother's history of skin cancer (yes=1 or no=0)} \\ x_{11} &= \text{mother's history of breast cancer (yes=1 or no=0)} \\ x_{12} &= \text{mother's history of diabetes (yes=1 or no=0) .} \end{aligned}$$

In this model most of the explanatory variables are dichotomous, i.e., they take on only one of two possible values, often taken to be 0 and 1, or -1 and 1. The linear model for these variables is

$$y_j = x_1\beta_{1,j} + \cdots + x_{12}\beta_{12,j} + e_j \qquad j = 1, ..., 7 \ .$$

If we have complete data on n persons (for which the values of all variables, including the y_j's, is known) we can estimate the unknown parameters $\beta_{i,j}$ (assuming n is sufficiently large) from the system (5.2.4) or (5.2.5) by least squares. For a data vector for the i-th person has the form

$$(x_{i,1},\, x_{i,2}, \ldots ,\, x_{i,12},\, y_{i,1},\, y_{i,2}, \ldots ,\, y_{i,7})$$

and allows us to construct the $n \times 12$ matrix X and the $n \times 7$ matrix Y used in (5.2.6).

5.2.25 Exercise. Check that the normal equations (5.2.6) are equivalent to m normal equations of the form (5.2.3) where β is taken to be $(b_{1j}, ..., b_{kj})'$ for some $j \in \{1, ..., m\}$. What computational advantage is there to using (5.2.6) rather than (5.2.3) m times?

5.2.26 Exercise. Consider an aircraft with position data (x_i, y_i, z_i) = (latitude, longitude, altitude) at times t_i, $i = 1, 2, ..., n$. Assume $t_1 < t_2 < \cdots < t_n$. Build a motion model for the aircraft based on second degree polynomials of the form $at^2 + bt + c$. Give the design matrix, the Y matrix, and the normal equations. We wish to use the model to accurately predict the aircraft's position at a time shortly after the last observation at t_n. Give a formula, based on this model, for the predicted position at $t_{n+1} > t_n$. Why are second degree polynomials a good choice for this application? What are the limitations of this model? What would happen if the difference $t_n - t_1$ is too large? too small?

5.2.27 Exercise. Write down a second order auto-regressive model for a time series y_k with vector values, $y_k \in \mathbb{R}^m$. What happens to the coefficients a_1 and a_2? What must the design matrix look like?

Infinite dimensional spaces

5.2.28 Example (polynomial approximation in L^2)**.** The Hilbert space in which the last few examples were set was $\mathbb{R}^n$ or $\mathbb{C}^n$. Now let a and b be two real numbers with $a < b$ and consider the Hilbert space $H = L^2((a,b);\mathbb{R})$. Suppose we are given a function $y(\cdot) \in H$ and we wish to find a polynomial

$$p(t) = \beta_1 + \beta_2 t + ... + \beta_k t^{k-1}$$

which best fits y in the sense that the real numbers β_j are chosen to minimize

$$\int_a^b (y(t) - p(t))^2 \, dt \ .$$

The functions t^j, $j = 0, 1, ..., k-1$, (which are of course vectors in H) are linearly independent (but we will not prove this here). A design 'matrix' for this problem is

$$X \ = \ [1 \ t \ t^2 \ \ldots \ t^{k-1}]$$

where the vectors t^j in H are thought of as column vectors. Using the inner product in H we form the $k \times k$ non-singular Gram matrix

$$X'X \ = \ \begin{pmatrix} \int_a^b 1 \, dt & \ldots & \int_a^b t^{k-1} \, dt \\ \vdots & & \vdots \\ \int_a^b t^{k-1} \, dt & \ldots & \int_a^b t^{k-1} t^{k-1} \, dt \end{pmatrix}$$

whose ij-th entry is $\int_a^b t^{i-1} t^{j-1} \, dt$, $i, j = 1, ..., k$, the inner product of t^{i-1} and t^{j-1} in H.

The least squares coefficients $\hat{\beta}_j$ satisfy the normal equations

$$X'X \, \hat{\beta} \ = \ X'y$$

where

$$X'y \ = \ \begin{pmatrix} \int_a^b y(t) \, dt \\ \vdots \\ \int_a^b t^{k-1} \, y(t) \, dt \end{pmatrix} \ .$$

5.2.29 Example (trigonometric approximation in L^2)**.** This approximation is precisely the Fourier series approximation to a square integrable function. We perform the same operations as in Example 5.2.28 but with sines and cosines replacing powers of t.

Let $H = L^2((0, 2\pi); \mathbb{R})$, fix $m \in \mathbb{N}$, and consider the trigonometric sum

$$g(t) \ = \ a_0 + \sum_{k=1}^{m} (a_k \cos kt + b_k \sin kt) \ . \tag{5.2.7}$$

(The set of functions $\cos kt$ and $\sin kt$, with $k \in \mathbb{N}_0$, is a complete set in H but this fact need not concern us here.) Given any $f \in H$ we want to choose the real numbers $a_0, ..., a_m, b_1, ..., b_m$ so that g is closest to f in the norm of H, that is, so that the expression

$$\int_0^{2\pi} |f(t) - g(t)|^2 \, dt$$

is minimized.

A schematic design 'matrix' for this problem is

$$X \ = \ \begin{pmatrix} 1 & \cos t & \sin t & \cos 2t & \sin 2t & \cdots & \cos mt & \sin mt \end{pmatrix}$$

where each function is thought of as a column vector. The Gram matrix is diagonal

$$(X'X) \;=\; \mathrm{diag}(2\pi, \pi, \pi, ..., \pi, \pi)$$

since $\int_0^{2\pi} \sin kt \;\sin \ell t\, dt = \int_0^{2\pi} \cos kt \;\cos \ell t\, dt = \pi\, \delta_{k\ell}$, and $\int_0^{2\pi} \sin kt \;\cos \ell t\, dt = 0$ for all k and ℓ. The inner product of the 'columns' of the design matrix with f gives

$$(X'f) = \Big(\int_0^{2\pi} 1\; f(t)\, dt, \int_0^{2\pi} \cos t\; f(t)\, dt, \int_0^{2\pi} \sin t\; f(t)\, dt, ..., \int_0^{2\pi} \cos mt\; f(t)\, dt, \int_0^{2\pi} \sin mt\; f(t)\, dt \Big)'.$$

The least squares formula for the coefficients in (5.2.7), $(X'X)^{-1}(Y'f)$, is precisely the classical formula for the Fourier coefficients of f

$$a_0 = \frac{1}{2\pi} \int_0^{2\pi} f(t)\, dt, \qquad a_k = \frac{1}{\pi} \int_0^{2\pi} \cos kt\, f(t)\, dt, \qquad b_k = \frac{1}{\pi} \int_0^{2\pi} \sin kt\, f(t)\, dt, \qquad k = 1, ..., m\ .$$

5.2.30 Example (processing a signal in noise)**.** In some applications of radar and sonar we desire to detect a 'signal', a linear combination of sinusoids of the form $e^{i\,\omega\, t}$, from data corrupted with 'noise'. The context here is that the electromagnetic or acoustic sensors are 'listening' only; there is no transmitted electromagnetic or acoustic wave, and no 'echo' to be detected. Such 'passive' systems can only detect objects that are radiating electromagnetic or acoustic waves; 'active' systems, which transmit a wave and watch for an echo, can detect objects that are not emitting any electromagnetic or acoustic energy.

Suppose we can observe a function $y(t)$ of time t for $0 < t < T$, and we believe that $y(t) = s(t) + n(t)$ is the sum of two other non-observable functions. The function $s(t)$ is a signal of the form

$$s(t) \;=\; \sum_{j=1}^{m} \alpha_j\, e^{i\omega_j t}$$

where the $\alpha_j \in \mathbb{C}$ are unknown amplitudes, and the ω_j's are real frequencies assumed known. The noise $n(t)$ models that part of the data $y(t)$ which is not represented by the signal model $s(t)$.

We estimate the α_j's by minimizing the norm of $n(t)$ in the Hilbert space $L^2(0,T)$. These estimates $\hat\alpha_j$ satisfy the normal equations

$$\begin{pmatrix} \int_0^T \overline{e^{i\omega_1 t}}\, e^{i\omega_1 t}\, dt & \cdots & \int_0^T \overline{e^{i\omega_1 t}}\, e^{i\omega_m t}\, dt \\ \vdots & & \vdots \\ \int_0^T \overline{e^{i\omega_m t}}\, e^{i\omega_1 t}\, dt & \cdots & \int_0^T \overline{e^{i\omega_m t}}\, e^{i\omega_m t}\, dt \end{pmatrix} \begin{pmatrix} \hat\alpha_1 \\ \vdots \\ \hat\alpha_m \end{pmatrix} = \begin{pmatrix} \int_0^T \overline{e^{i\omega_1 t}}\, y(t)\, dt \\ \vdots \\ \int_0^T \overline{e^{i\omega_m t}}\, y(t)\, dt \end{pmatrix}. \tag{5.2.8}$$

The next example is the basis for many applications in electrical engineering.

5.2.31 Example (Hilbert space of random variables)**.** Let Ω be a sample space with probability P, and let $L^2(\Omega, P)$ be the Hilbert space of real valued random variables on Ω which have finite variance.[43] Suppose m real random variables $x_1, x_2, ..., x_m$ are given (perhaps observable in some experiment) and it is required to estimate the value of another real random variable y from the values of the x's. If the correlations (second order moments) among all random variables $y, x_1, ..., x_m$ are known we can find the best linear predictor, $a_1x_1 + \cdots + a_mx_m$, of y by choosing the constants $a_1, ..., a_m$ to minimize the expected value

$$\mathbb{E}[(y - a_1x_1 - \cdots - a_mx_m)^2],$$

that is, the error in the Hilbert space $L^2(\Omega, P)$. We know that the error random variable $y - a_1x_1 - \cdots - a_mx_m$ must be orthogonal to each of the x_i's:

$$\mathbb{E}[x_i(y - a_1x_1 - \cdots - a_mx_m)] \;=\; 0 \qquad\qquad i = 1, ..., m\ .$$

[43] P is a measure on a σ-field of measurable subsets of Ω, but these technicalities need not concern us. The reader need only be familiar with elementary properties of probability and random variables, and to accept the fact (from the theory of the Lebesgue integral) that the vector space of random variables with finite variance is complete.

Using the linearity of the integral $\mathbb{E}(\cdot)$ this gives

$$\mathbb{E}(x_i y) \;=\; a_1\mathbb{E}(x_i x_1) + \cdots + a_m\mathbb{E}(x_i x_m) \qquad\qquad i = 1, ..., m\ . \tag{5.2.9}$$

Therefore the vector $a = (a_1, ..., a_m)'$ satisfies the matrix equation $Ra = b$ where R is the $m \times m$ matrix with ij-th entry $\mathbb{E}(x_i x_j)$ and b is the column vector with i-th component $\mathbb{E}(x_i y)$. The system (5.2.9) is precisely the normal equations in the Hilbert space of random variables $L^2(\Omega, P)$.

5.2.32 Exercise. Find the quadratic $at^2 + bt + c$ which best approximates $\sin(t)$ on $0 < t < \pi$ in the norm of $L^2(0, \pi)$. Find the cubic $at^3 + bt^2 + ct + d$ which best approximates $\cos(t)$ on $0 < t < \pi$ in the norm of $L^2(0, \pi)$.

5.2.33 Exercise. Write a computer program to numerically evaluate the matrices $X'X$ and $X'y$ in Example 5.2.28 when a, b, k, and $y(t)$ are input. If possible, include code to solve for $\hat{\beta}$ and to plot both $y(t)$ and the least squares polynomial $p(t)$ on the interval (a, b). Pick various examples for $y(t)$ and k and plot the results.

5.2.34 Exercise (polynomial approximation in L^2)**.** Let $\Omega \subset \mathbb{R}^2$ be a bounded open set. Suppose it is required to approximate the function e^{x+iy} by a second degree polynomial $b_{00} + b_{10}x + b_{01}y + b_{20}x^2 + b_{11}xy + b_{02}y^2$ for $(x, y) \in \Omega$. Give a formula for the complex coefficients b_{kl} that provide the best approximation in the norm of $L^2(\Omega; \mathbb{C})$.

5.2.35 Exercise (approximation of $f(x)$ and $f'(x)$)**.** Consider the Hilbert space $H^1(0, 1)$ of $L^2(0, 1)$ functions whose derivatives are also in $L^2(0, 1)$. The inner product here is $\langle f, g\rangle_1 = \int_0^1 fg + f'g'\,dx$. Suppose we want to approximate both $f(x)$ and $f'(x)$ on $(0, 1)$ by an m-th degree polynomial $p(x) = \sum_0^m a_j x^j$. Set up the normal equations for this problem.

5.2.36 Exercise. Generalize Exercise 5.2.35 to more variables and higher order derivatives. Give the required Hilbert space and sketch what the normal equations would look like.

5.2.37 Exercise (using covariance for prediction)**.** A certain university desires to predict calculus scores for students so that remedial and supportive measures can be taken before failure occurs. It is decided that the key variables (at least those for which data is available) are $x_1 =$ economic privilege, $x_2 =$ parents' education, and $x_3 =$ score in pre-calculus, as well as $y =$ calculus score (of course). Historical data shows that the covariance matrix for these variables is

$$\mathrm{Cov}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ y \end{pmatrix} \;=\; \begin{pmatrix} 1 & 0 & 0.5 & 0.25 \\ 0 & 1 & 0.5 & 0.25 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.25 & 0.25 & 0.5 & 1 \end{pmatrix}$$

when all are scaled to have mean zero and values between -5 and 5. A student who wants to take calculus has the following personal data: $x_1 = -2$, $x_2 = 1$, and $x_3 = -1$. Predict his calculus score. Is there anything suspicious about the covariance data given here? Speculate on a model which adds the variable $x_4 =$ number of hours per day spent gaming. What kind of covariance values would you expect?

5.2.38 Exercise. Work out the normal equations in Example 5.2.31 when all random variables are allowed to be complex valued.

5.2.39 Exercise. In this exercise we extend Example 5.2.31 to vector valued random variables. Consider the Hilbert space $H = L^2(\Omega, P) \otimes \mathbb{R}^n$ with inner product $\langle x, y\rangle = \mathbb{E}(x'y)$ where x and y are column vectors. Let $x_1, ..., x_m$, and y be elements of H. Find the vector of constants $\beta = (\beta_1, ..., \beta_m)' \in \mathbb{R}^m$ which minimizes

$$\mathbb{E}|y - X\beta|^2 \;=\; \mathbb{E}|y - x_1\beta_1 - \cdots - x_m\beta_m|^2$$

where $X = [x_1, ..., x_m]$ is the $n \times m$ matrix of random variables. In particular, show the Gram matrix can be written in the form $\mathbb{E}(X'X)$ and give its individual components. Give the normal equations. Formulate and solve this problem with $\mathbb{R}^n$ replaced by $\mathbb{C}^n$.

5.2.40 Exercise (conditional density with respect to sub-algebra of events)**.** Let $P(x)$ be a non-decreasing differentiable function on $[0,1]$ with $P(0)=0$ and $P(1)=1$. Let $p(x)=P'(x)$, so P is a cumulative probability distribution on $[0,1]$ and p its density. Let $n\in\mathbb{N}$ and, for $i=1,...,n$, $I_i=(\frac{i-1}{n},\frac{i}{n})$ be a partition of the interval $(0,1)$. For the purposes of numerical computation it is often desirable to approximate $p(x)$ using the indicator functions $1_{I_i}(x)$. Find the projection of $p(x)$ onto the span of these indicator functions in $L^2(0,1)$. This is the *conditional density of* $p(x)$ *with respect to the sub-algebra* generated by the events $\{I_i=(\frac{i-1}{n},\frac{i}{n})\ ;\ i=1,2,...,n\}$ in the algebra of all events in the sample space $[0,1]$.

5.2.41 Exercise. Generalize Example 5.2.34 to the case that $y(t)$ is a row vector valued function with m components, and $H=L^2((a,b);\mathbb{R}^m)$. If $x(t)$ and $y(t)$ are m-vector functions the inner product in H is

$$\langle x,y\rangle \;=\; \int_a^b x(t)\cdot y(t)\,dt \;=\; \int_a^b x_1(t)y_1(t)+\cdots+x_m(t)y_m(t)\,dt\ .$$

The approximating polynomial $p(t)$ in Example 5.2.34 is also vector valued with each β_j a row vector in $\mathbb{R}^m$.

Regularized estimates and ridge regression

Non-linear least squares estimates

5.3 Consistency of Linear Least Squares Estimates

If $y_1,y_2,...,y_n$ is a random sample from any distribution with mean μ and finite variance σ^2, then the sample mean $\overline{y}=\frac{1}{n}\sum_1^n y_j$ has expected value $\mathbb{E}\overline{y}=\mu$ and, using independence, variance $\mathbb{E}(\overline{y}-\mathbb{E}(\overline{y}))^2=\sigma^2/n$. This implies that the 'mean squared error' of $\overline{y}$, as an estimate of μ, is

$$\mathbb{E}(\overline{y}-\mu)^2=\sigma^2/n\to 0 \qquad \text{as} \qquad n\to 0.$$

We say that $\overline{y}$ is (mean square) *consistant* for the parameter μ. We also notice that the rate of 'convergence' of $\overline{y}$ to μ, as the sample size n increases, is of order $1/n$. This rate of convergence for parametric statistical models is typical, and we will now show that it holds more generally for least squares estimates under natural assumptions.

We now assume the data $\{(x_i,y_i);i=1,...,n\}$ is available, where $x_i\in\mathbb{R}^d$ and $y\in\mathbb{R}$. We are interested in a model of the form

$$y \;=\; \beta_1\,h_1(x)+\cdots+\beta_p\,h_p(x)+error \tag{5.3.1}$$

where *error* is a random error, the β's are unknown parameters, and the functions $h_j(x)$ are known. Taking h to be a monomial in the components of x is a common example.

The least squares estimates of the model parameters are given as follows. We build the $n\times p$ deterministic design matrix

$$X \;=\; \begin{pmatrix} h_1(x_1) & \cdots & h_p(x_1)\\ \vdots & \vdots & \vdots\\ h_1(x_n) & \cdots & h_p(x_n)\end{pmatrix}_{n\times p} \tag{5.3.2}$$

which we assume is full rank, with $n\geq p$. The model can then be written $y=X\beta+e$ with components $y_i=\beta_1h_1(x_i)+\cdots+\beta_ph_p(x_i)+e_i$. The least squares estimate for $\beta=(\beta_1,...,\beta_p)'$ is

$$\hat{\beta} \;=\; (X'X)^{-1}X'y\,. \tag{5.3.3}$$

If we assume $e=(e_1,...,e_n)'$ has mean zero and covariance matrix σ^2I, then

$$E\hat{\beta} \;=\; \beta \qquad \text{and} \qquad \mathrm{Cov}(\hat{\beta}) \;=\; \sigma^2\,(X'X)^{-1}\,. \tag{5.3.4}$$

Let's consider the matrix $X'X$; its $k\ell$-th entry is $\sum_{i=1}^n h_k(x_i)\,h_\ell(x_i)$. As $n \to \infty$ this expression will in general grow without bound. But it is reasonable to expect that its average, $\frac{1}{n}\sum_{i=1}^n h_k(x_i)\,h_\ell(x_i)$, may have a definite limiting value. For instance when $d = 1$, $p = 2$, $h_1(x) = 1$ (constant), and $h_2(x) = x$, we have

$$X'X \;=\; \begin{pmatrix} n & \sum_1^n x_i \\ \sum_1^n x_i & \sum_1^n x_i^2 \end{pmatrix} \qquad \text{and} \qquad \tfrac{1}{n}X'X \;=\; \begin{pmatrix} 1 & \overline{x} \\ \overline{x} & s^2 \end{pmatrix}$$

where $\overline{x}$ denotes the sample average and s^2 denotes the sample second moment (dependence on n is suppressed).

5.3.1 Theorem. *Suppose* $\lim_{n\to\infty} \frac{1}{n}X'X$ *exists in* $\mathbb{R}^{p\times p}$ *and that this limit has non-zero determinant. Then for any* $a \in \mathbb{R}^p$ *the (scalar) estimate* $a'\hat{\beta}$ *of* $a'\beta$ *is mean square consistant,* $E(a'\hat{\beta} - a'\beta)^2 \to 0$ *as* $n \to \infty$*, when the data* y *satisfies the model* $y = X\beta + e$ *with* $e \sim (0, \sigma^2 I)$*.*

Proof. From $y = X\beta + e$, $e \sim (0, \sigma^2 I)$, and (5.3.3) we have

$$\begin{aligned} E(a'\hat{\beta} - a'\beta)^2 \;&=\; E\{a'(X'X)^{-1}X'e\,e'X(X'X)^{-1}a\} \\ &=\; \sigma^2\,a'(X'X)^{-1}a \;=\; \tfrac{\sigma^2}{n}\,a'(\tfrac{1}{n}X'X)^{-1}a \;\to\; 0 \end{aligned} \tag{5.3.5}$$

as $n \to \infty$. □

5.3.2 Theorem. *Suppose there is a compact set* $K \subset \mathbb{R}^{p\times p}$ *which lies inside the open complement of* $\det^{-1}(0)$*, the pull-back of* $\{0\}$ *by the determinant* $\det : \mathbb{R}^{p\times p} \to \mathbb{R}$*, and an* $n_0 \in \mathbb{N}$ *such that* $\frac{1}{n}X'X$ *lies inside* K *whenever* $n \geq n_0$*. Then for any* $a \in \mathbb{R}^p$ *the (scalar) estimate* $a'\hat{\beta}$ *of* $a'\beta$ *is mean square consistant,* $E(a'\hat{\beta} - a'\beta)^2$ *as* $n \to \infty$*, when the data* y *satisfies the model* $y = X\beta + e$ *with* $e \sim (0, \sigma^2 I)$*.*

Proof. For a fixed $a \in \mathbb{R}^p$, the mapping $f : \mathbb{R}^{p\times p} \to \mathbb{R}$, given by $f(A) = a'A^{-1}a$, is continuous on the complement of $\det^{-1}(0)$. Since K is compact, $f(K)$ is a bounded subset of $\mathbb{R}$. Hence the sequence $f(\frac{1}{n}X'X)$ is bounded when $n \geq n_0$. The computations in (5.3.5) hold, so we conclude that

$$E(a'\hat{\beta} - a'\beta)^2 \;=\; \tfrac{\sigma^2}{n}\,a'(\tfrac{1}{n}X'X)^{-1}a \;=\; \tfrac{\sigma^2}{n}\,f(\tfrac{1}{n}X'X) \;\to\; 0$$

as $n \to \infty$. □

5.3.3 Corollary. *Under the assumptions of the preceding theorem,* $E\,(\hat{\beta}_i - \beta_i)^2 \to 0$ *for each* $i = 1, ..., p$*.*

5.3.4 Remark. It is also true, under very reasonable assumptions, that $\hat{\beta}$ is asymptotically normal. This fact would allow us to do hypothesis testing, for instance, to ask if the climate really is warming up (is the slope > 0)?

5.3.5 Remark. Mean square consistency implies some weaker types of consistency, as well.

6 The Fourier Transform

6.1 Definition of the Fourier Transform

The *Fourier transform* of a function $f : \mathbb{R} \to \mathbb{C}$ is defined as

$$\hat{f}(\omega) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega x}\, f(x)\, dx \tag{6.1.1}$$

where $\omega \in \mathbb{R}$, when the (improper) integral exists. In particular, when $f \in L^1(\mathbb{R})$ the Fourier transform exists because for any fixed ω,

$$|\hat{f}(\omega)| \;\le\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |e^{-i\omega x}\, f(x)|\, dx \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |f(x)|\, dx \;<\; \infty\,.$$

The value $\hat{f}(\omega)$ gives the amount the frequency ω contributes to the function $f(x)$. This is seen by the *Fourier inversion* formula:

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\omega x}\, \hat{f}(\omega)\, d\omega \tag{6.1.2}$$

valid, again, when the integral exists. The inversion formula will be proved later.

For clairity let's decompose the real and imaginary parts of the Fourier transform. Recall that $e^{-i\omega x} = \cos\omega x - i\sin\omega x$, and let $f(x) = u(x) + iv(x)$ where u and v are real valued. Then

$$\hat{f}(\omega) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (u(x)\cos\omega x + v(x)\sin\omega x)\, dx + \frac{i}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (v(x)\cos\omega x - u(x)\sin\omega x)\, dx\,.$$

Notice that the Fourier transform is linear. When f and g are appropriate functions and a and b are complex numbers,

$$\widehat{(af+bg)}(\omega) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\omega x} \big(af(x) + bg(x)\big)\, dx \;=\; a\hat{f}(\omega) + b\hat{g}(\omega)\,.$$

6.1.1 Example. If $a > 0$, the Fourier transform of the symmetric step function $1_{[-a,a]}(x)$ is

$$\frac{1}{\sqrt{2\pi}} \int_{-a}^{a} e^{-i\omega x}\, dx = \frac{1}{\sqrt{2\pi}} \Big[\frac{e^{-i\omega x}}{-i\omega}\Big]_{x=-a}^{a} = \frac{1}{\sqrt{2\pi}} \Big[\frac{e^{+i\omega a} - e^{-i\omega a}}{i\omega}\Big] = \sqrt{\frac{2}{\pi}} \Big[\frac{\sin\omega a}{\omega}\Big] = \sqrt{\tfrac{2}{\pi}}\, a\, \operatorname{sinc} a\omega\,.$$

6.2 Basic Properties of the Fourier Transform

In order to use the Fourier transform to solve useful problems we make use of some of its interesting properties. Before listing these in a theorem we need the following definition.

6.2.1 Definition. Whenever the following integral exists, the *convolution* of two functions f and g is defined by the formula

$$f{*}g(x) \;=\; \int_{-\infty}^{\infty} f(x-y)g(y)\, dy\,.$$

The convolution is a new function, denoted $f{*}g$, defined on the real line.

6.2.2 Example. Assume, wothout any loss in generality, that $b - a \le d - c$, i.e., that the interval $[a, b]$ is shorter than $[c, d]$. Then the convolution of the two 'box car' functions $f(x) = 1_{[a,b]}(x)$ and $g(x) = 1_{[c,d]}(x)$ is

$$f{*}g(x) = \int_c^d 1_{[a,b]}(x-y)\, dy = \begin{cases} 0 & \text{if} \quad x < a+c \\ x - (a+c) & \text{if} \quad a+c < x < a+d \\ b-a & \text{if} \quad b+c < x < a+d \\ (b+d) - x & \text{if} \quad a+d < x < b+d \\ 0 & \text{if} \quad b+d < x \end{cases}.$$

The reader should diagram the overlapping rectangle and its changing area as x increases.

The change of variables $z = x - y$ in the definition shows that $f*g(x) = \int f(z)g(x-z)\,dz$ as well. Therefore, $f*g = g*f$; the convolution is commutative.

6.2.3 Theorem. *Let $f \in L^1(\mathbb{R})$ be complex valued, and let $s \in \mathbb{R}$ and $t > 0$.*

(i) If $g(x) = e^{isx}f(x)$ then $\hat{g}(\omega) = \hat{f}(\omega - s)$.

(ii) If $g(x) = f(x - s)$ then $\hat{g}(\omega) = e^{-is\omega}\hat{f}(\omega)$.

(iii) If $g \in L^1(\mathbb{R})$ and $h = f*g$, then $\hat{h}(\omega) = \sqrt{2\pi}\,\hat{f}(\omega)\,\hat{g}(\omega)$.

(iv) If $g(x) = \overline{f(-x)}$ then $\hat{g}(\omega) = \overline{\hat{f}(\omega)}$.

(v) If $g(x) = f(x/t)$ then $\hat{g}(\omega) = t\hat{f}(t\omega)$.

(vi) If $g(x) = -ixf(x)$ and $g \in L^1(\mathbb{R})$, then $\hat{f}$ is differentiable and $\hat{g}(\omega) = \hat{f}'(\omega)$.

(vii) If $g(x) = f'(x) \in L^1(\mathbb{R})$ then $\hat{g}(\omega) = i\omega\hat{f}(\omega)$.

(viii) If all derivatives of $f(x)$ of orders $\le k$ exist and belong to $L^1(\mathbb{R})$, and if $g(x) = f^{(k)}(x)$, then $\hat{g}(\omega) = (i\omega)^k\hat{f}(\omega)$.

Proof. (i) Compute

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega x}\,g(x)\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega x}e^{isx}f(x)\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i(\omega-s)x}\,f(x)\,dx = \hat{f}(\omega - s).$$

(ii) Compute

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega x}\,g(x)\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega x}f(x-s)\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega(y+s)}\,f(y)\,dy = e^{-i\omega s}\hat{f}(\omega),$$

where we have made the substitution $y = x - s$.

(iii) Compute, using Fubini's theorem,

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega x}\Big(\int_{-\infty}^{\infty} f(x-y)g(y)\,dy\Big)\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\Big(\int_{-\infty}^{\infty} e^{-i\omega x}f(x-y)g(y)\,dx\Big)\,dy\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\Big(\int_{-\infty}^{\infty} e^{-i\omega(x-y)}\,e^{-i\omega y}f(x-y)g(y)\,dx\Big)\,dy\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\Big(\int_{-\infty}^{\infty} e^{-i\omega(x-y)}f(x-y)\,dx\Big)e^{-i\omega y}g(y)\,dy = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\Big(\sqrt{2\pi}\hat{f}(\omega)\Big)e^{-i\omega y}g(y)\,dy\\
&= \sqrt{2\pi}\hat{f}(\omega)\hat{g}(\omega)\,.
\end{aligned}$$

Here we have used the fact that $\int_{-\infty}^{\infty} u(x-y)\,dx = \int_{-\infty}^{\infty} u(x)\,dx$ whenever $u \in L^1(\mathbb{R})$ and $y \in \mathbb{R}$.

(iv) Use the change of variables $y = -x$ to compute

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega x}\overline{f(-x)}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{+i\omega y}\overline{f(y)}\,dy = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\overline{e^{-i\omega y}f(y)}\,dy.$$

(v) Use the change of variables $y = x/t$ to compute

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega x}f(x/t)\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega yt}f(y)\,t\,dy = t\hat{f}(t\omega).$$

(vi) The proof that $\hat{f}$ is differentiable is most easily accomplished by applying the Lesbegue dominate convergence theorem to the difference quotient. We will omit this part and only give the formal verification that $\hat{g}(\omega) = \hat{f}'(\omega)$.

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega x}(-ix)f(x)\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{d\,e^{-i\omega x}}{d\omega}f(x)\,dx = \frac{1}{\sqrt{2\pi}}\frac{d}{d\omega}\int_{-\infty}^{\infty} e^{-i\omega x}f(x)\,dx = \hat{f}'(\omega).$$

We have here assumed, without proof, that we can interchange the order of differentiation and integration.

(vii) We will only give a formal derivation of this formula, assuming that the following integration by parts is valid:

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega x} f'(x)\,dx = \Big[e^{-i\omega x} f(x)\Big]_{x=-\infty}^{\infty} - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \frac{d\,e^{-i\omega x}}{dx} f(x)\,dx = \frac{i\omega}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-i\omega x} f(x)\,dx,$$

where we have used the fact that $\lim_{x\to\pm\infty} f(x) = 0$ since $f \in L^1(\mathbb{R})$. The formula (vii) is easily remembered by differentiating the Fourier inversion formula, and passing the $\frac{d}{dx}$ under the integral sign. We have not used this as a proof since we have not yet proven this formula.

(viii) This formula is obtain by interating formula (vii). □

6.2.4 Exercise. Make use of Example 6.1.1 and Theorem 6.2.3(i), to compute the Fourier transform of $1_{[a,b]}(t)$ for any finite interval $[a, b]$. You may want to use the new parameters $c = (a + b)/2$ and $d = b - c$.

6.2.5 Exercise. Assuming the Fourier transform of y and its derivatives exist, use Theorem 6.2.3(vii) to take the transform of both sides of the homogeneous differential equation

$$a_m\, y^{(m)} + a_{m-1}\, y^{(m-1)} + \cdots + a_1\, y' + a_0\, y \;=\; 0$$

to obtain the algebraic equation

$$\Big(a_m(i\omega)^m + a_{m-1}(i\omega)^{m-1} + \cdots + a_1 i\omega + a_0\Big)\,\hat{y}(\omega) \;=\; 0\,.$$

Here the a_k's are constants. How would you now find solutions of this differential equation?

We now state carefully a useful condition when the Fourier inversion formula holds.

6.2.6 Theorem. *There is a bijection (one-to-one, onto function) $\mathcal{F} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$, called the* Fourier transform*, with the following properties.*

(a) When $f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, $\mathcal{F}(f) = \hat{f}$ is defined by the right side of (6.1.1).

(b) When $f \in L^2(\mathbb{R})$, $\|\mathcal{F}(f)\|_0 = \|f\|_0$ so that $\mathcal{F}$ is an isometry of $L^2(\mathbb{R})$ onto itself.

(c) For any $f \in L^2(\mathbb{R})$ and $R > 0$, the cut-off function $f_R(t) = f(t)1_{[-R,R]}(t)$ belongs to $L^1(\mathbb{R})$, and $\mathcal{F}(f_R) \to \mathcal{F}(f)$ in the $L^2(\mathbb{R})$ norm as $r \to \infty$.

(d) If for any $\hat{f} \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ we define $\mathcal{F}^{-1}(\hat{f}) = f$ as in (6.1.2), then **FINISH THIS STEVE**

The L^2 function $\mathcal{F}(f)$ is usually denoted $\hat{f}$.

Proof. The (rather long) proof may be found in Rudin, *Real and Complex Analysis*, Theorem 9.13. □

Fourier transform on $\mathbb{R}^d$ The definition and properties of the Fourier transform are also valid for functions of d variables. The defining formulas are

$$\hat{f}(\omega) \;=\; \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} e^{-i\omega\cdot x}\, f(x)\,dx \tag{6.2.1}$$

and

$$f(x) \;=\; \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} e^{i\omega\cdot x}\, \hat{f}(\omega)\,d\omega \tag{6.2.2}$$

where $\omega \cdot x = \sum_1^d \omega_j x_j$.

All points in Theorem 6.2.3 are valid when x and ω are in $\mathbb{R}^d$ provided statements are given the appropriate (obvious) vector interpretation, except that $\sqrt{2\pi}$ is replaced by $(2\pi)^{d/2}$, the right side of (v) becomes $t^d\hat{f}(t\omega)$ with t still a scalar, and k in (viii) is now a multi-index.

6.3 Some Applications of the Fourier Transform

Solutions of Certain Linear Constant Coefficient Differential Equations We can use Theorem 6.2.3 to turn linear constant coefficient differential equations into algebraic equations. In order to solve for the solution function we have to invert its Fourier transform. By 'Certain' we mean the integrals involving $1/P(\xi)$, where P is a polynomial must exist; in critical applications they sometimes do not, and the method presented in this section is not useful.

6.3.1 Example (Solve the PDE $u - \Delta u = f(x)$ on $\mathbb{R}^d$)**.** Let $f \in L^2(\mathbb{R}^d)$ and suppose we want to solve $u - \Delta u = f(x)$ for a function $u(x)$, perhaps in $L^2(\mathbb{R}^d)$ or a 'better' function space. Proceeding formally, we take the Fourier transform of this equation to obtain

$$\hat{u}(\xi) - [\Delta u]\hat{}\,(\xi) = \hat{u}(\xi) - (-\xi_1^2 - \xi_2^2 - \cdots - \xi_d^2)\hat{u}(\xi) \;=\; (1 + |\xi|^2)\hat{u}(\xi) \;=\; \hat{f}(\xi)\,.$$

The function $1 + |\xi|^2$ is bounded away from 0 on $\mathbb{R}^d$ so we may divide by it to obtain

$$\hat{u}(\xi) \;=\; \frac{\hat{f}(\xi)}{1 + |\xi|^2}$$

which is certainly in $L^2(\mathbb{R}^d)$ if f is (Theorem 6.2.6(b)). . We may give the following formula for the solution:

$$f(x) \;=\; \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{\hat{f}(\xi)}{1 + |\xi|^2} e^{i\xi\cdot x}\, d\xi\,.$$

The improper integral exists as a limit in $L^2(\mathbb{R}^d)$. In fact, by Schwarz inequality in $L^2(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} \Big|\frac{\hat{f}(\xi)}{1 + |\xi|^2}\Big|\, d\xi \;\le \Big\{ \int_{\mathbb{R}^d} \frac{1}{(1 + |\xi|^2)^2}\, d\xi \;\int_{\mathbb{R}^d} |\hat{f}(\xi)|^2\, d\xi \Big\}^{1/2}$$

whenever the first integral on the right side is finite. In this case $\hat{f}/(1 + |\xi|^2) \in L^1(\mathbb{R}^d)$ and we can be sure that $f(x)$ is continuous.

6.3.2 Exercise. Proceeding formally as we have done above, and assuming the Fourier transform of the δ-function $\delta(x)$ is $(2\pi)^{-d/2}$, a constant function on $\mathbb{R}^d$, determine how large $k \in \mathbb{N}$ must be to ensure the solution of

$$(I - \Delta)^k u(x) \;=\; \delta(x)\,, \qquad x \in \mathbb{R}^d\,,$$

is continuous.

Hint. Recall that $u \in C(\mathbb{R}^d)$ if $\hat{u} \in L^1(\mathbb{R}^d)$. Also, one can integrate $\int (1 + |\xi|^2)^{-k}\, d\xi$ over $\mathbb{R}^d$ by switching to polar coordinates: $d\xi = r^{d-1} dr\, d\sigma(\theta)$ where $r = |\xi| > 0$, $\theta = (\theta_1, ..., \theta_{d-1})$, and $d\sigma$ is the surface area element on the surface of the unit sphere in $\mathbb{R}^d$.

6.3.3 Exercise. Try to use the method in Example 6.3.1 to solve Laplace's equation $\Delta u = f$ on $\mathbb{R}^d$. What goes wrong? (Use polar coordinates to study the integral $\int (1/|\xi|^2)\, d\xi$ on $\mathbb{R}^d$.)

6.3.4 Example (Solve the PDE $u_{tt} - \Delta u = 0$ on $\mathbb{R}^{d+1}$)**.** We wish to solve the equation $u_{tt} - \Delta u = 0$ for $u = u(x,t)$, where $x \in \mathbb{R}^d$ and $t \in \mathbb{R}$, and with $u(x,0) = f(x)$ and $u_t(x,0) = g(x)$ given functions. Proceeding formally as in the last example, we consider the Fourier transform in the space variables alone

$$\hat{u}(\xi, t) \;=\; \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} u(x,t) e^{-i\xi\cdot x}\, dx\,.$$

Applying this transform to our initial value problem we obtain

$$\hat{u}_{tt} + |\xi|^2 \hat{u} \;=\; 0 \qquad \text{and} \qquad \hat{u}(\xi,0) = \hat{f}(\xi)\ ,\ \hat{u}_t(\xi,0) = \hat{g}(\xi)\,.$$

For each fixed ξ we solve this ODE and obtain

$$\hat{u}(\xi,t) \;=\; A(\xi)\, \cos|\xi|t \;+\; B(\xi)\, \sin|\xi|t,$$

where A and B are functions we determine from the initial conditions: $A(\xi) = \hat{f}(\xi)$ and $B(\xi) = \hat{g}(\xi)/|\xi|$. Thus,

$$\hat{u}(\xi, t) \;=\; \hat{f}(\xi)\, \cos|\xi|t + \hat{g}(\xi)\, \frac{\sin|\xi|t}{|\xi|}$$

and we can give the formula

$$u(x,t) \;=\; \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \{\hat{f}(\xi)\, \cos|\xi|t + \hat{g}(\xi)\, \frac{\sin|\xi|t}{|\xi|}\}\, e^{i\xi\cdot x}\, d\xi\,.$$

Notice that the function $(\sin|\xi|t)/|\xi|$ is actually bounded for $\xi \in \mathbb{R}^d$.

6.3.5 Exercise. Proceeding formally as we have been, solve the initial value problem $u_{tt} - \Delta u = f(x,t)$ on $\mathbb{R}^{d+1}$ where f is a nice function and the solution satisfies $u(x,0) = 0$ and $u_t(x,0) = 0$ on $\mathbb{R}^d$.

6.3.6 Exercise. Proceeding formally, solve the 'telegraph equation' $u_{tt} + r\, u_t - \Delta u = 0$ on $\mathbb{R}^{d+1}$ where $r > 0$ is a resistance and the initial 'signal' is given by $u(x,0) = f(x)$ and $u_t(x,0) = g(x)$ on $\mathbb{R}^d$.

6.3.7 Exercise. In our first course in differential equations we learn that we can find solutions of a linear constant coefficient equation

$$a_m y^{(m)} + a_{m-1} y^{(m-1)} + \cdots + a_1 y' + a_0 y \;=\; 0$$

by plugging in functions of the form $y(x) = e^{rx}$ and finding solutions r of the algebraic equation $a_m r^m + \cdots + a_1 r + a_0 = 0$. This tactic also works for linear constant coefficient partial differential equations, equations of the form

$$\sum_{|\alpha|\le m} a_\alpha\, \partial^\alpha u(x) \;=\; 0\,, \tag{6.3.1}$$

where $a_\alpha \in \mathbb{C}$. Show that any function of the form $u(x) = e^{\zeta\cdot x}$, where $\zeta \in \mathbb{C}^d$ and $\zeta \cdot x = \sum_1^d \zeta_j x_j$, will be a solution of this equation if ζ is a root of the polynomial equation

$$\sum_{|\alpha|\le m} a_\alpha\, \zeta^\alpha \;=\; 0\,. \tag{6.3.2}$$

Suppose you could find several such roots, $\zeta^{(1)}, ..., \zeta^{(n)}$, of (6.3.2). How could you then form more general solutions of (6.3.1)?

Consider the wave equation in Example 6.3.4, $u_{tt} - \Delta u = 0$, $x \in \mathbb{R}^d$ (without initial conditions), and possible solutions of the form $u(x,t) = e^{i\omega t} e^{i\eta\cdot x}$. Find the algebraic relation that couples ω and $\eta = (\eta_1, ..., \eta_d)$. Sketch the set of solutions in the $\omega\eta$-plane when $d = 1$. What does this picture look like (in $\mathbb{R}^3$) when $d = 2$?

Nyquist Sampling Rate In modern electrical engineering systems the first step in signal processing is very often digitizing an analog signal. The question naturally arises of what digital sample rate is required if one needs to process signals in a certain frequency range $[-\omega_0, \omega_0]$. An ideal answer to this question is given by the Nyquist sampling theorem. In practice (because real-life signals and A/D converters are noisy) it is usually necessary to digitize at 1.5 to 2 or more times the Nyquist rate.

Bandpass Filters Suppose the function $x(t)$ is a signal, that is, a real valued function of time. Very often x is one's voice, a radio broadcast, or a television program which has been translated into electromagnetic waves for transmission through the atmosphere. We assume that x is 'bandlimited' which is to say that it has a representation of the form $x(t) = \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} e^{i\omega t}\, \hat{x}(\omega)\, d\omega$ for some bounded function $\hat{x}$ with support in the bounded interval $-a < \omega < a$.

Electrical engineers face the problem of broadcasting numerous signals simultaneously without interference. One way to do this is by amplitude modulation, and a simple highly idealized version of this method can be explained in the following way. Multiply the signal $x(t)$ by a 'carrier' $e^{i\omega_0 t}$ and transmit the new amplitude modulated signal

$$y(t) \;=\; x(t)\, e^{i\omega_0 t}$$

in place of x. Using Theorem 6.2.3 (i), the Fourier transform of y is

$$\hat{y}(\omega) \;=\; \hat{x}(\omega - \omega_0)\,.$$

This is the Fourier transform of the original signal shifted by the carrier frequency ω_0, and has support in the interval $(\omega_0 - a, \omega_0 + a)$. By choosing different carrier frequencies, $\omega_1, \omega_2, ..., \omega_n$ (different channels), each separated from its neighbor by at least $2a$, many different signals can be simultaneously transmitted without interfering with each other. The sum of all simultaneously transmitted signals is

$$y(t) \;=\; \sum_1^n y_j(t) \;=\; \sum_1^n x_j(t)\, e^{i\omega_j t}$$

where $x_j(t)$ is the j-th signal and n may be quite large.

When a device (radio, TV, cell phone) receives all the signals being transmitted, it can pick out the one of interest using a bandpass filter. Let $W(\omega_0, a; \omega) = \mathbf{1}_{(\omega_0 - a, \omega_0 + a)}(\omega)$ and $w_{\omega_0, a}(t)$ be the inverse Fourier transform of $W(\omega_0, a; \cdot)$. When the receiving device is 'tuned' to the k-th frequency (channel) it filters the received signal $y(t)$ by applying the formula

$$w_{\omega_k, a} * y(t) \;=\; \int w_{\omega_k, a}(t - s)\, y(s)\, ds \qquad \text{or} \qquad W(\omega_k, a; \omega)\, \hat{y}(\omega)$$

which removes all unwanted signals leaving only $y_k(t)$ or $\hat{y}_k(\omega)$. The carrier function $e^{i\omega_k t}$ can then be eliminated by division and the signal $x_k(t)$ (e.g., Beethoven's 5th Symphony) recovered.

I NEED A COUPLE DIAGRAMS TO ILLUSTRATE.

Causal Filters

Central Limit Theorem