



연세대학교
YONSEI UNIVERSITY



수학계산학부(계산과학공학)
School of Mathematics and Computing
(Computational Science and Engineering)

Quantum Partial Differential Equations

Jeongbin Jo

jeongbin033@yonsei.ac.kr

Department of Physics, Yonsei University
School of Mathematics and Computing (Computational Science and Engineering),
Yonsei University, Seoul, 03722, Republic of Korea



October 8, 2026

Abstract

This report develops quantum algorithms for partial differential equations along a single pipeline: a continuous PDE is discretized into a linear system or a semidiscrete evolution, the resulting matrix is block-encoded, and a quantum transformation—quantum singular value transformation, Hamiltonian simulation, or a quantum linear-system solver—is applied before a quantity of interest is read out. Linear-system solvers sit at the center of that pipeline, and inversion is the Cauchy calculus for $f(z) = 1/z$: HHL, QSVT, and contour quadratures of resolvents are three writings of one identity. Semigroups and Hamiltonian simulation are the neighboring maps e^{tz} and e^{-itz} . Classical PDE theory, the Cauchy theory of the plane, and numerical discretization form the first part; fault-tolerant quantum linear algebra the second; and elliptic, hyperbolic, parabolic, and nonlinear applications the third.

Contents

I	Classical Analysis and Numerical Discretization	1
1	Motivation	2
1.1	What the equations are used for	2
1.2	Why they are expensive	3
1.3	What a quantum computer changes	4
1.4	What it does not change	4
1.5	Where an advantage can appear	5
1.6	What the rest of the report is for	6
2	Introduction to Partial Differential Equations	7
2.1	Ordinary initial-value problems	7
2.2	Partial Differential Equations	9
2.3	Initial and boundary conditions	12
2.4	Well-posedness	12
2.5	Types of second-order equations	13
2.6	Analyticity and the Laplace equation	17
3	Analytic solution formulas	19
3.1	The heat kernel	19
3.2	The Newtonian potential	25
3.3	Green's function	28
3.4	The wave equation	36
3.5	Series solutions and special functions	41
3.6	Nonlinear equations	45
4	The Cauchy theory	51
4.1	Line integrals	51
4.2	Goursat's theorem	51
4.3	Primitives in a disc	52
4.4	Two real integrals	54
4.5	Cauchy's integral formulas	55
4.6	Liouville's theorem and the identity theorem	57
4.7	Further consequences	57
4.8	Harmonic conjugates and the Poisson kernel	60
4.9	Conformal maps	62

5	Discretization and the linear-algebra reduction	64
5.1	Preliminaries	64
5.2	Rootfinding	66
5.3	Interpolation	70
5.4	Polynomial approximation	71
5.5	Quadrature	73
5.6	Grid functions and stencils	74
5.7	Dirichlet Poisson: how the matrix is assembled	76
5.8	Neumann conditions: assembling the matrix	80
5.9	Periodic conditions, circulants, and the DFT	82
5.10	Which transform for which boundary	84
5.11	Finite elements, briefly	84
5.12	Method of lines and time stepping	85
5.13	Consistency, stability, and convergence	87
5.14	Conditioning, sparsity, and what a quantum algorithm sees	89
II	Fault-Tolerant Quantum Linear Algebra and Primitives	91
6	Block encoding and qubitization	92
6.1	Block encodings	92
6.2	An algebra of block encodings	93
6.3	Linear combination of unitaries	94
6.4	Sparse-matrix and finite-difference block encodings	95
6.5	Qubitization	98
7	Quantum signal processing and singular value transformation	102
7.1	Quantum signal processing	102
7.2	Quantum singular value transformation	105
7.3	Example targets: inverse, exponential, and oscillatory filters	108
8	Hamiltonian simulation	113
8.1	Trotter–Suzuki product formulae	114
8.2	Linear combination of unitaries and truncated Taylor series	117
8.3	Qubitization and QSP-based simulation	118
8.4	Time-dependent simulation and the interaction picture	120
8.5	From Hermitian simulation to non-Hermitian generators	122
9	Solving linear systems of equations	125
9.1	The quantum linear systems problem	125
9.2	Classical connection	126
9.3	Quantum phase estimation and the HHL algorithm	126
9.4	QSVT-based linear-system solver	127
9.5	Cauchy’s formula as the common writing	127
9.6	Worked example: Poisson’s equation	127
9.7	Near-optimal and optimal solvers	128
9.8	What this chapter does not cover	128

10 Solving linear differential equations	129
10.1 Problem setup	129
10.2 The linear-systems method	129
10.3 Time marching	129
10.4 Linear combination of Hamiltonian simulation	129
10.5 Example: the heat equation	129
10.6 Non-Hermitian generators and complex absorbing potentials	129
11 Complex analysis as a language for PDEs and quantum algorithms	130
11.1 The identity	130
11.2 Three levels of the interpretation	131
11.3 What is already known	131
11.4 A dictionary	132
11.5 From Cauchy–Riemann to elliptic PDEs	132
11.6 Resolvents, Green’s functions, and linear systems	132
11.7 Semigroups, Hamiltonian simulation, and LCHS	132
11.8 Residues, projectors, and what QPE is counting	133
11.9 QSVT as a real-spectrum surrogate	133
III Quantum Algorithms for Differential Equations	134
12 Quantum algorithms for elliptic PDEs	135
12.1 Classical discretization and the direct QLSP route	135
12.2 First-order factorization and Hermitian dilation	135
12.3 QFT-based spectral filtering	135
12.4 Elliptic eigenvalue problems and phase estimation	135
13 Quantum algorithms for hyperbolic PDEs	136
13.1 Classical wave discretizations	136
13.2 From the wave equation to a Schrödinger equation	136
13.3 Schrödingerization	136
13.4 LCHS for non-Hermitian and forced hyperbolic systems	136
14 Quantum algorithms for parabolic PDEs	137
14.1 Classical heat discretizations	137
14.2 Crank–Nicolson as a quantum linear system	137
14.3 QSVT implementation of the heat semigroup	137
14.4 Hermitian dilation and Gaussian transmutation	137
15 Nonlinear PDEs	138
15.1 Why nonlinear PDEs require a different viewpoint	138
15.2 From nonlinear PDEs to nonlinear ODEs	138
15.3 Carleman linearization	138
15.4 Kolmogorov, Liouville, and Koopman–von Neumann formulations	138
15.5 Classical rootfinding as a supporting tool	138

IV	NISQ Implementations and Research Frontiers	139
16	Variational linear solvers and NISQ methods	140
16.1	The VQLS cost function	140
16.2	Relation to the QLSP	140
16.3	Limitations and when a variational method is not a solver	140
17	Latest research and outlook	141
17.1	Recent algorithmic developments	141
17.2	Software and programming interfaces	141
17.3	Open questions	141

Part I

Classical Analysis and Numerical Discretization

Chapter 1

Motivation

A quantum computer does not solve a partial differential equation by inventing a new closed-form solution. It stores a discretized unknown in the amplitudes of a state and applies a matrix function—an inverse, a semigroup, or a unitary group—to that state [25, 34, 17]. Whether this is cheaper than a classical solver depends on the size of the problem, on how well conditioned the discretized operator is, and on what one is allowed to read out. This chapter describes those dependences qualitatively. The cost accounting is carried out once the matrices exist, in Chapters 5 and 9.

1.1 What the equations are used for

Differential equations are the models of quantities that change in space or in time. The same formal object covers the electric potential in a device, the temperature in a solid, the pressure and velocity in a fluid, a plasma, an acoustic or electromagnetic wave, the electrons of a molecule, the price of an option, the currents in an electrical network, and the size of a population [30, 17]. In each case the unknown is a function, the equation constrains its derivatives, and a useful answer is rarely the formula: it is a number extracted from the solution, such as a flux through a surface, a total energy, a voltage at a terminal, or the value of a payoff.

Steady states are elliptic. The electrostatic potential in a semiconductor device or a capacitor solves a Poisson equation whose source is the charge density; the temperature in a solid at equilibrium solves Laplace’s or Poisson’s equation; the displacement of a loaded bridge or airframe solves the equations of linear elasticity, an elliptic system [30, 33]. The engineer rarely wants the whole field. A capacitance, the heat flux through one face, or the stress at a critical joint is a single linear functional of the solution.

Diffusion is parabolic. Heat conduction and the spreading of a chemical species smooth their initial data as time increases. Finance and statistics produce the same equation with a different meaning. The Black–Scholes equation for the price of an option on a single asset becomes the heat equation after a change of variables [11], and the Fokker–Planck equation evolves the probability density of a stochastic process [44]. In both, the number of space variables is the number of random factors, so a basket option on d assets, or a stochastic model with d state variables, is a d -dimensional PDE. The wanted output is an expectation: a price, a mean, or the probability of reaching a region.

Propagation without smoothing is hyperbolic. Acoustic and elastic waves carry the signals of ultrasound and seismic imaging, and Maxwell’s equations govern radar scattering, antennas, and optical devices. Energy is conserved and data travel at finite speed; the output is a signal recorded

at a receiver or the energy scattered in one direction. The Schrödinger equation is not hyperbolic, but it shares the conservative character that matters here: its evolution is unitary, which is also the native operation of a quantum computer. For a molecule it carries three coordinates per electron, which makes it the standard example of a high-dimensional equation, and its low-lying states are what quantum chemistry and drug design ask for [17].

Many of the most consequential models are nonlinear. The Navier–Stokes equations describe the air around a wing and the atmosphere in a weather forecast. The Vlasov equation describes a plasma in a six-dimensional phase space of positions and velocities, which is one reason fusion-device design is expensive. Reaction–diffusion systems describe combustion, chemical patterns, and the spread of a population or an epidemic in space [17]. Here the difficulty is not only size: products of the unknown with itself break the linear structure that the quantum primitives of this report require, and Section 1.5 returns to what can be done about it.

Some models of the same kind are not PDEs at all. The currents in an electrical network obey a differential-algebraic system whose size is the number of components, and circuit simulators such as SPICE solve such systems at industrial scale [17]. Viscoelastic materials and epidemics with an incubation period have memory: the rate of change depends on an integral of the past solution. A spatially discretized PDE is itself a large system of ordinary differential equations, so these models and the PDEs meet in the same linear algebra.

equation (type)	use	quantity of interest	function of A
Poisson, Laplace, linear elasticity (elliptic)	device potentials, steady heat, loaded structures	capacitance, heat flux, stress at a point	A^{-1}
heat, Black–Scholes, Fokker–Planck (parabolic)	diffusion, option pricing, stochastic dynamics	mean, price, probability of an event	e^{tA} , decaying
wave, Maxwell (hyperbolic); Schrödinger	acoustics, seismic and radar imaging, quantum chemistry	signal at a receiver, scattered energy, energy levels	e^{-itH} , unitary
Navier–Stokes, Vlasov, reaction–diffusion (nonlinear)	weather, aerodynamics, fusion plasmas, epidemics	lift and drag, heat load, concentration	linear only after a lift
RLC networks, equations with memory	circuit simulation, viscoelasticity, epidemics	terminal voltage, strain, case count	A^{-1} or e^{tA} after a lift

Table 1.1: Where the equations of this report come from, what is read out of them, and which function of the discretized matrix A (or Hamiltonian H) a quantum algorithm has to apply.

Table 1.1 collects these cases. Two features recur and drive the rest of this chapter. The answer that matters is a few derived numbers, not the whole field. And the problems that matter are large: fine three-dimensional grids, or models with many coordinates. The last column anticipates the report: after discretization, each class of equation asks for a different function of one matrix.

1.2 Why they are expensive

The dimension of the unknown is the first obstruction. A grid with n nodes along each side of a d -dimensional domain has

$$N = n^d = \Theta(h^{-d}) \quad (1.1)$$

degrees of freedom when the mesh width is $h = 1/n$. Refining the mesh, or adding a coordinate (a species, an asset, a momentum variable), multiplies N . The high-dimensional models

of Section 1.1—Fokker–Planck equations, multi-asset pricing, kinetic plasma models, molecular Schrödinger equations—and fine three-dimensional grids are expensive for this reason, before any question of quantum hardware arises [33, 17].

After discretization, a linear elliptic problem is a sparse system $A\mathbf{u} = \mathbf{f}$ with N unknowns, and a linear evolution is a system of N ordinary differential equations $\mathbf{u}' = A\mathbf{u}$. Classical solvers for these systems are mature. A multigrid cycle, or a fast transform on a simple geometry, returns the whole vector in work nearly proportional to N : one pays essentially for writing the answer down [33]. Storage follows the same count, since the classical representation of $\mathbf{u}$ is a list of N numbers. In high dimension the bottleneck is N itself, not the solver.

A second quantity enters through the mesh. For a second-order operator the condition number $\kappa(A)$ grows like the inverse square of the mesh width (equation (5.78)), so every refinement makes the linear system harder to solve. Iterative methods without a preconditioner slow down accordingly; multigrid and related preconditioners are the classical devices that remove this dependence. The same conditioning reappears in every quantum solver.

Producing every nodal value cannot take fewer than N operations, on a quantum device or a classical one. Any asymptotic improvement has to be an improvement about a *coarser* output: a state, an inner product, or one derived quantity of interest [25, 34].

1.3 What a quantum computer changes

The quantum representation of the same vector is the normalized state

$$|u\rangle = \frac{1}{\|\mathbf{u}\|} \sum_{j=0}^{N-1} u_j |j\rangle \quad (1.2)$$

on $\lceil \log_2 N \rceil$ qubits. The index j is a computational-basis label for a grid node; the value u_j is an amplitude. The register grows with the logarithm of the number of nodes. A three-dimensional grid with 10^3 nodes on a side has $N = 10^9$ entries and needs 30 qubits for the node register; a ten-dimensional grid with n nodes on a side needs $10 \log_2 n$ qubits while $N = n^{10}$ is already unstorable. Ancillas for the block encoding and for the transformation add further small registers, not a register of size N .

Two classical tasks become two quantum tasks [17, 35]. A linear solve $A\mathbf{u} = \mathbf{f}$ becomes the preparation of a state proportional to $A^{-1}|f\rangle$. A linear evolution $\mathbf{u}' = A\mathbf{u}$ becomes the preparation of a state proportional to $e^{tA}|u_0\rangle$, which is a Hamiltonian simulation when A is skew-Hermitian and a semigroup when A is dissipative. In both cases the algorithm implements a function of one matrix. The later chapters separate those functions; they do not describe different physical theories.

Given efficient access to a sparse or block-encoded A and to the state $|f\rangle$, the Harrow–Hassidim–Lloyd algorithm and its successors prepare $A^{-1}|f\rangle$ with a number of queries that grows only polylogarithmically in N , and polynomially in κ and in $1/\varepsilon$ for a target accuracy ε [25, 35]. Set against a classical cost that grows at least like N , this is the origin of the phrase “exponential speedup”. The precise statements are in Chapter 9.

1.4 What it does not change

The polylogarithmic dependence on N holds in the query model. It survives end to end only under conditions that sit outside the linear-algebra step [1, 34]. Each condition below is either a classical–quantum interface or a property of the discretization.

- *Loading.* If the right-hand side is an arbitrary list of N numbers, a circuit that prepares $|f\rangle$ needs a number of gates that grows with N , even though the state it prepares is logarithmically wide. Smooth data, a short Fourier series, or a tensor-product initial condition can be loaded cheaply.
- *Matrix access.* A stencil with a bounded number s of nonzeros per row has an efficient sparse-access encoding. A dense unstructured matrix does not [35, 34].
- *Readout.* The output is a state, not a list. Recovering every entry of $\mathbf{u}$ by tomography repeats the whole algorithm at least on the order of N times, and the improvement disappears. What can stay cheap is a single observable—a voltage, a flux, an overlap $\langle c|u\rangle$ —estimated by repeated measurement, with more repetitions for a tighter accuracy.
- *Normalization.* The state (1.2) forgets $\|\mathbf{u}\|$. A physical quantity that depends on the magnitude of the solution needs that norm estimated separately.
- *Conditioning.* The query count grows with κ , and refining the mesh to reduce the discretization error raises κ . Solver tolerance and mesh width are therefore one budget, which is why the condition number is treated as a consequence of the discretization (Section 5.13.5) rather than as a free parameter of the circuit. Preconditioning matters for a quantum solver for the same reason it matters for a classical one.
- *Time and dissipation.* For an evolution the role of κ is played by the evolution time multiplied by the norm of the generator, which also grows as the mesh is refined. When A is dissipative the solution decays, and a state proportional to a small final vector costs more to prepare [6].

None of these is removed by a faster inverter in the middle of the pipeline [34, 17].

1.5 Where an advantage can appear

The fair comparison is not between a quantum algorithm that returns one observable and a classical algorithm that writes out the whole field. It is between two algorithms that return the same quantity to the same accuracy, with the mesh chosen for that accuracy on both sides. Montanaro and Pallister carry out this comparison for elliptic problems discretized by finite elements, and Linden, Montanaro, and Shao for the heat equation [42, 36]. In both analyses the quantum advantage is at most polynomial, and it becomes significant only as the spatial dimension grows. Low-dimensional Poisson problems, for which fast classical solvers already work in time close to N , are the wrong place to look. High-dimensional linear problems, and structured network problems whose size N is the number of components rather than a number of grid nodes, are the natural candidates [17, 34]. The quantitative version of this comparison is deferred to the model Poisson problem of Section 9.6.

Krovi describes the same target in an interview on IBM’s research programme: discretize, map to a linear system or a Hamiltonian simulation, and encode the vector in amplitudes [17]. For linear RLC networks, which are differential-algebraic equations, the interview reports a quantum procedure whose cost grows logarithmically with the number of components, where a conventional circuit simulator grows at least linearly. The conditions of Section 1.4 apply to that claim as to any other. The equations with memory of Section 1.1 are lifted to a larger ordinary differential equation and treated by the same primitive.

Nonlinear equations need one more step. Carleman linearization lifts a dissipative nonlinear system to a linear one in a larger space, and the lift is provably efficient when dissipation dominates the nonlinearity [37]. The Navier–Stokes equations are accessible in this way only in a weakly

nonlinear, laminar regime. A turbulent flow, at large Reynolds number, is outside that regime, and no current quantum reduction removes that fact [17]. The lift is the subject of Chapter 15; it is not a reason to ignore the linear cost model.

1.6 What the rest of the report is for

The cost model uses only a few quantities: the number of unknowns N and the spatial dimension d , the sparsity s , the condition number $\kappa(h)$, the solver tolerance ε , the evolution time for time-dependent problems, and the choice of observable. Each of them is fixed by the continuous problem and by its discretization. The type of the PDE decides whether the matrix function is an inverse, a decaying exponential, or an oscillation. The boundary conditions decide whether A is invertible and which nodes are unknown. The mesh decides κ . None of these is visible if the discussion starts from a circuit.

The chapters that follow therefore stay on one reduction. Chapter 2 records the continuous equation, its type, and the data that make it well-posed. Chapter 3 writes the solution as a convolution against a kernel. Chapter 4 is the complex calculus of the two-dimensional Laplacian: Cauchy's theorem, the statements equivalent to it, and the conformal maps that move the disk. Chapter 5 turns that kernel into a sparse matrix and computes the s and κ that the discussion above refers to. The quantum chapters implement functions of that matrix. Elliptic, parabolic, hyperbolic, and nonlinear problems reappear as choices of the function and of the matrix, not as separate theories of computation.

Chapter 2

Introduction to Partial Differential Equations

A differential equation relates an unknown function to its derivatives. If that unknown depends on a single independent variable, the equation is ordinary: $F(x, u, u', u'', \dots) = 0$, as in $u' = f(x, u)$ or $u'' + u = 0$. Section 2.1 records the linear initial-value theory. The partial case begins when the unknown depends on two or more independent variables and the equation involves partial derivatives [30]. In two variables a first-order PDE has the form $F(x, y, u, u_x, u_y) = 0$, often written $F(x, y, u, p, q) = 0$ with $p = u_x$ and $q = u_y$; a second-order PDE is $F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = 0$. The same writing in n variables is $F(x, u, Du, D^2u, \dots) = 0$ on a domain in $\mathbb{R}^n$.

The aim is to treat these equations as one class of objects—an unknown u , an operator L , and data that select a solution—rather than as three unrelated methods. Linearity, characteristics, and the sign of a discriminant distinguish cases; they do not split the subject. This chapter records those distinctions and the data that make a problem well-posed.

2.1 Ordinary initial-value problems

An ordinary equation determines a function of one independent variable. For a linear initial-value problem the solution is an explicit formula, and that formula is the scalar case of the matrix functions named in Chapter 1 [32, 7].

The initial-value problem asks for a differentiable function u satisfying

$$u'(x) = f(x, u), \quad u(x_0) = u_0, \quad (2.1)$$

where f is continuous for every (x, u) in a domain D of the plane. A solution on an interval $[a, b]$ containing x_0 is a function whose graph lies in D , which takes the value u_0 at x_0 , and whose derivative equals $f(x, u(x))$ at every point of the interval.

For the linear equation

$$u' = \lambda u + g(x), \quad u(0) = u_0, \quad (2.2)$$

with g continuous on $[0, \infty)$, an integrating factor produces the solution in closed form. Multiplying (2.2) by $e^{-\lambda x}$ gives $\frac{d}{dx}(ue^{-\lambda x}) = e^{-\lambda x}g(x)$. Integrating from 0 to x and multiplying through by $e^{\lambda x}$ yields Duhamel's formula

$$u(x) = u_0 e^{\lambda x} + \int_0^x e^{\lambda(x-t)} g(t) dt. \quad (2.3)$$

The first term propagates the initial value by the exponential semigroup $e^{\lambda x}$. The integral propagates the forcing by the same semigroup. A negative real λ makes the exponential decay, and a

purely imaginary λ makes it a rotation. These are the scalar forms of the factors e^{tA} and e^{-itH} in Table 1.1.

The same formula applies to a second-order linear equation once it is written as a first-order system. In

$$u'' = a_1(x)u' + a_0(x)u + g(x), \quad u(x_0) = \alpha, \quad u'(x_0) = \beta, \quad (2.4)$$

set $u_1 = u$ and $u_2 = u'$. The pair (u_1, u_2) satisfies

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}' = \begin{pmatrix} 0 & 1 \\ a_0(x) & a_1(x) \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + \begin{pmatrix} 0 \\ g(x) \end{pmatrix}, \quad \begin{pmatrix} u_1(x_0) \\ u_2(x_0) \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}. \quad (2.5)$$

Duhamel's formula for (2.5) replaces the scalar factor $e^{\lambda(x-t)}$ by the fundamental matrix of the coefficient matrix. When the coefficients are constant, that fundamental matrix is a matrix exponential. A spatially discretized evolution is this constant-coefficient case with a large matrix in place of the companion matrix, and that exponential is the function the later algorithms apply. The passage from a partial differential equation to this system is carried out in Section 5.12.

For a general f , existence and uniqueness hold on a short interval. Let D be the rectangle $|x - x_0| < \alpha$, $|u - u_0| < \beta$, let f be continuous on D and Lipschitz continuous in u ,

$$|f(x, u_1) - f(x, u_2)| \leq K|u_1 - u_2| \quad (2.6)$$

for some $K \geq 0$ and all $(x, u_1), (x, u_2) \in D$, and let M be a bound for $|f|$ on D . Set $\gamma = \min\{\alpha, \beta/M\}$. Then (2.1) has a unique solution on $[x_0 - \gamma, x_0 + \gamma]$ [32, 7]. For (2.2) the Lipschitz constant is $|\lambda|$, and (2.3) supplies the solution for every $x \geq 0$.

The solution depends continuously on the data. For the perturbed problem $u' = f(x, u) + \delta(x)$ with $u(x_0) = u_0 + \varepsilon$ and δ continuous, every sufficiently small perturbation admits a unique solution $u(x; \delta, \varepsilon)$ on the same interval. Whenever $\gamma K < 1$,

$$\max_{|x-x_0| \leq \gamma} |u(x) - u(x; \delta, \varepsilon)| \leq \frac{1}{1 - \gamma K} (|\varepsilon| + \gamma \|\delta\|_\infty). \quad (2.7)$$

Existence, uniqueness, and (2.7) are well-posedness for the initial-value problem. Section 2.4 asks the same three conditions of a partial differential equation.

The constant in (2.7) can be large on a longer interval. The linear problem

$$u' = 100u - 101e^{-x}, \quad u(0) = 1 \quad (2.8)$$

has the exact solution $u(x) = e^{-x}$. The same equation with initial value $1 + \varepsilon$ has the exact solution $e^{-x} + \varepsilon e^{100x}$. On any fixed interval the two solutions differ by at most a constant times $|\varepsilon|$, so the problem remains well-posed, and the constant grows like e^{100x} . The exponent is the Lipschitz constant $K = 100$ of the right-hand side, the same K that enters (2.7) and the factor $e^{\lambda x}$ in (2.3). After the unknown is replaced by a grid vector, this amplification is the conditioning described in Section 1.4. Figure 2.1 plots the two solutions for $\varepsilon = 10^{-6}$.

A quadratic right-hand side has a different closed form. The equation $u' = -u^2$ has the equilibrium solution $u \equiv 0$ and the further solutions $u(x) = 1/(x + c)$ for an arbitrary constant c . Such an equation becomes an instance of (2.3) only after a transformation that restores linearity, taken up in Chapter 15.

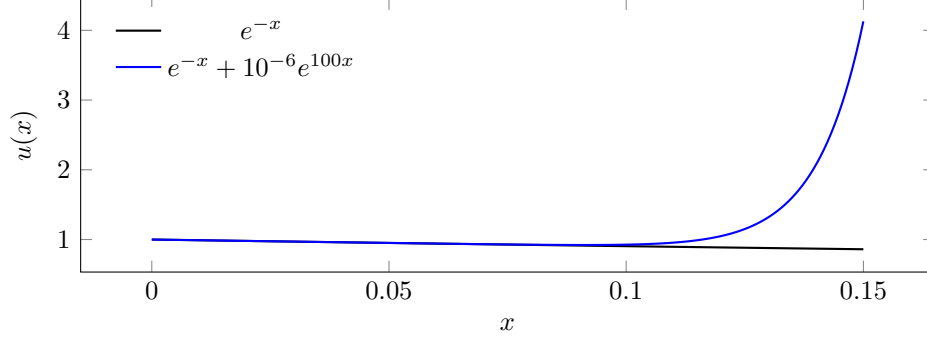


Figure 2.1: The initial-value problem (2.8) and the same equation with initial value $1 + 10^{-6}$. The solutions start together and then separate by the factor e^{100x} .

2.2 Partial Differential Equations

Writing L for the differential operator that sends u to the left-hand side of $L(u) = 0$, the equation is *linear* when $L(\alpha u_1 + \beta u_2) = \alpha L(u_1) + \beta L(u_2)$ for all scalars α, β [30]. If u_1 and u_2 both solve $L(u) = 0$, so does every linear combination. The transport operator $L(u) = u_t + cu_x$ is linear in this sense. If the identity fails, the equation is nonlinear: for the shock-wave (inviscid Burgers) operator $L(u) = u_t + uu_x$ one has $L(u_1 + u_2) = (u_1 + u_2)_t + (u_1 + u_2)(u_1 + u_2)_x \neq 0$ in general, even when $L(u_1) = L(u_2) = 0$.

A first-order linear example is the transport equation $u_t + cu_x = 0$. The representative second-order linear equations are

$$\Delta u = 0 \quad (\text{Laplace}), \quad u_t = k\Delta u \quad (\text{heat}), \quad u_{tt} = c^2\Delta u \quad (\text{wave}). \quad (2.9)$$

They are the constant-coefficient normal forms of the elliptic, parabolic, and hyperbolic types in Section 2.5.

2.2.1 First-order linear equations

The constant-coefficient equation

$$au_x + bu_y = 0 \quad (2.10)$$

says that the directional derivative of u along (a, b) vanishes. Level curves $u = \text{const}$ are orthogonal to $\nabla u = (u_x, u_y)$, so (a, b) is tangent to those curves: $a dx + b dy = 0$, hence $ax + by = \text{const}$. The general solution is therefore an arbitrary profile of the complementary invariant [30],

$$u(x, y) = f(-bx + ay). \quad (2.11)$$

The same conclusion follows from a change of variables. Set $\xi = ax + by$ and $\eta = -bx + ay$, which is invertible when $a^2 + b^2 \neq 0$. The chain rule gives $u_x = au_\xi - bu_\eta$ and $u_y = bu_\xi + au_\eta$, so

$$au_x + bu_y = a(au_\xi - bu_\eta) + b(bu_\xi + au_\eta) = (a^2 + b^2)u_\xi. \quad (2.12)$$

Thus (2.10) becomes $u_\xi = 0$, and u depends only on η : again $u = f(-bx + ay)$.

As an initial-value illustration, $2u_t + 3u_x = 0$ with $u(x, 0) = \sin x$ has general solution $u(x, t) = f(3t - 2x)$. The initial datum forces $f(-2x) = \sin x$, so $f(s) = \sin(-s/2)$ and $u(x, t) = \sin(x - 3t/2)$ [30]. The transport equation $u_t + cu_x = 0$ is the same geometry with speed c : solutions are rightward shifts $u(x, t) = f(x - ct)$ of the initial profile.

When the coefficients depend on (x, y) ,

$$a(x, y)u_x + b(x, y)u_y = 0 \quad (2.13)$$

is $\mathbf{v} \cdot \nabla u = 0$ with $\mathbf{v} = (a, b)$. Along the integral curves of $\mathbf{v}$,

$$\frac{d}{ds}u(x(s), y(s)) = 0, \quad \frac{dx}{ds} = a(x, y), \quad \frac{dy}{ds} = b(x, y), \quad (2.14)$$

so u is constant on those *characteristic curves*. For $a = y$ and $b = -x$ the characteristic system is $dx/ds = y$, $dy/ds = -x$, hence $x^2 + y^2 = \text{const}$ and $u(x, y) = f(x^2 + y^2)$. Figure 2.2 draws those circles. The same characteristic geometry, applied to the principal symbol, later distinguishes elliptic, parabolic, and hyperbolic second-order equations: it decides how data propagate.

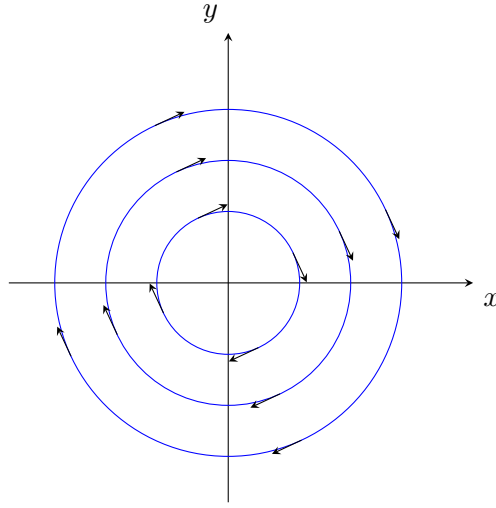


Figure 2.2: Characteristic curves of $yu_x - xu_y = 0$. The solution is constant on each circle $x^2 + y^2 = \text{const}$, and the arrows are the direction $(y, -x)$.

2.2.2 First-order nonlinear equations

A model nonlinear transport equation is inviscid Burgers' equation

$$u_t + uu_x = 0. \quad (2.15)$$

If u_1 and u_2 both solve (2.15), their sum generally does not [30]:

$$L(u_1 + u_2) = (u_1 + u_2)_t + (u_1 + u_2)(u_1 + u_2)_x = L(u_1) + L(u_2) + u_1(u_2)_x + u_2(u_1)_x, \quad (2.16)$$

and the cross terms need not vanish. Characteristics may intersect in finite time, producing shocks. For the initial datum $u(x, 0) = 1 - x$ the characteristic through ξ is the line $x = \xi + t(1 - \xi)$, and these lines meet at $(1, 1)$, as drawn in Figure 2.3. Superposition is lost: nonlinearity is not a small perturbation of (2.10), but a change in the algebraic type of the equation.

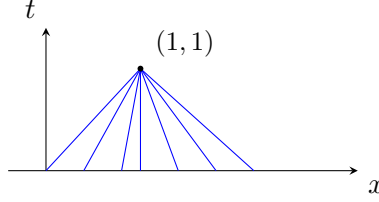


Figure 2.3: Characteristics of inviscid Burgers' equation (2.15) for $u(x, 0) = 1 - x$. The line through ξ has speed $1 - \xi$, and the family meets at $(x, t) = (1, 1)$.

2.2.3 Second-order linear equations

The models in (2.9) arise from elementary balance laws [30]. The first-order transport equation is already a conservation statement: if a density u on $[0, a]$ is merely shifted by speed c , then $\int_0^a u(x, t) dx = \int_{ch}^{a+ch} u(x, t + h) dx$ for $h \neq 0$. Differentiating first in a and then in h at $h = 0$ recovers $u_t + cu_x = 0$.

All three models below follow one template. Let u be the density of a conserved quantity on $\mathbb{R}^n$ and J its flux, the amount crossing a unit area per unit time. For an arbitrary region D with outward normal ν , the rate of change of the content equals the inflow, and the divergence theorem turns the boundary integral into a volume integral:

$$\frac{d}{dt} \int_D u dx = - \int_{\partial D} J \cdot \nu dS = - \int_D \nabla \cdot J dx \implies u_t + \nabla \cdot J = 0. \quad (2.17)$$

The region is arbitrary, so the integrands agree pointwise. The equation of motion is fixed by the constitutive law that expresses J through u . A drift at speed c is $J = cu$ and gives transport. Heat flows from hot to cold in proportion to the gradient, $J = -k\nabla u$ (Fourier's law, or Fick's law for a dissolved substance), and gives $u_t = k\Delta u$. For a fluid the flux of mass is ρv , which gives the continuity equation (3.169) of Section 3.6.3. The same law $J = -k\nabla u$ also shows why Laplace's equation is the stationary case: $u_t = 0$ means $\nabla \cdot J = 0$, a flux with no net source.

For the heat equation, let $u(x, t)$ be the temperature on a wire. Conservation of heat on an arbitrary interval $[a, b]$, together with Fourier's law that the flux is $-ku_x$, gives

$$\frac{d}{dt} \int_a^b u dx = ku_x(b, t) - ku_x(a, t) = \int_a^b ku_{xx} dx, \quad (2.18)$$

hence $u_t = ku_{xx}$. In $\mathbb{R}^d$ the same argument with the divergence theorem yields $u_t = k\Delta u$. Stationary temperature ($u_t = 0$) is Laplace's equation $\Delta u = 0$; solutions are called harmonic functions.

The one-dimensional wave equation is the small-slope limit of a vibrating string. Let $u(x, t)$ be the transverse displacement, ρ the linear density, and T the tension, which is constant when horizontal motion is neglected. At a point the string has the unit tangent $(1, u_x)/\sqrt{1 + u_x^2}$, so the vertical force that the rest of the string exerts on the piece over $[a, b]$ is $T[u_x/\sqrt{1 + u_x^2}]_a^b$. Newton's law for that piece reads

$$\rho \int_a^b u_{tt} dx = T \left[\frac{u_x}{\sqrt{1 + u_x^2}} \right]_a^b = \int_a^b \partial_x \left(\frac{T u_x}{\sqrt{1 + u_x^2}} \right) dx. \quad (2.19)$$

Since a and b are arbitrary, $\rho u_{tt} = T \partial_x (u_x/\sqrt{1 + u_x^2})$, which is nonlinear. For small slopes the Taylor expansion $(1 + s)^{-1/2} = 1 - \frac{s}{2} + \dots$ gives $u_x/\sqrt{1 + u_x^2} = u_x - \frac{1}{2}u_x^3 + \dots$, and keeping the first term yields $u_{tt} = c^2 u_{xx}$ with $c = \sqrt{T/\rho}$. On a membrane with surface density ρ and tension T

per unit length, the vertical force on the piece over a region $D \subset \mathbb{R}^2$ is, to first order in the slope, $T \oint_{\partial D} \partial_\nu u \, dS = T \int_D \Delta u \, dx$. Thus $\rho \int_D u_{tt} = T \int_D \Delta u$ for every D , and $u_{tt} = c^2 \Delta u$ with $c = \sqrt{T/\rho}$.

The same three operators therefore appear as balance laws, not as three separate theories: a conservation statement for transport, Fourier's law for heat (and its stationary limit Laplace), and Newton's law for a string or membrane.

2.2.4 Second-order nonlinear equations

Second-order nonlinear equations appear as soon as the flux or the reaction depends on u itself: viscous Burgers $u_t + uu_x = \nu u_{xx}$, reaction–diffusion $u_t = \Delta u + f(u)$, or geometric equations such as the minimal surface equation. The classification of the next sections applies to the linear principal part of such an equation; the lower-order nonlinear terms do not change the type.

2.3 Initial and boundary conditions

A differential equation by itself does not select a unique solution. Data on the domain, or on its boundary, are required, and the type of the equation dictates which data are natural [30].

For a parabolic equation such as the heat equation one prescribes an *initial condition*

$$u(x, 0) = \phi(x). \quad (2.20)$$

A hyperbolic equation such as the wave equation is second order in time, so both position and velocity are given:

$$u(x, 0) = \phi(x), \quad u_t(x, 0) = \psi(x). \quad (2.21)$$

An elliptic equation (Laplace, Poisson) has no distinguished time variable; the unknown is determined by boundary values on $\partial\Omega$.

On a bounded spatial domain Ω the standard boundary conditions are:

- *Dirichlet*: $u|_{\partial\Omega} = g$;
- *Neumann*: $\partial_n u|_{\partial\Omega} = h$;
- *Robin*: $(\alpha u + \beta \partial_n u)|_{\partial\Omega} = f$.

Periodic conditions on a torus, or decay at infinity on $\mathbb{R}^d$ (or a half-space), play the same role: they are boundary conditions at infinity [30]. Neumann data for Poisson's equation are compatible only if the forcing has zero mean, which is the integral of $-\Delta u = f$ over Ω . Choosing data that the type does not admit produces an ill-posed problem, not a slightly weaker one.

2.4 Well-posedness

Following Hadamard, a PDE problem is *well-posed* when a solution exists, is unique, and depends continuously on the data [30]. Stability means, roughly, that the solution is controlled by the data; if any of the three fails, the problem is ill-posed. The ordinary problem of Section 2.1 meets these three requirements when f is Lipschitz continuous in u . The constant in (2.7) may be large, as in (2.8), and the problem is still well-posed. The example below is a failure of continuous dependence.

A standard illustration is the Cauchy problem for Laplace's equation in the upper half-plane $\mathbb{R}_+^2 = \{(x, y) : y > 0\}$. The functions

$$u_n(x, y) = \frac{1}{n} e^{-\sqrt{n}} \sin(nx) \sinh(ny) \quad (2.22)$$

are harmonic, vanish at $y = 0$, and have $\partial_y u_n(x, 0) = e^{-\sqrt{n}} \sin(nx) \rightarrow 0$ as $n \rightarrow \infty$. At the interior point $(\pi/2, 1)$, with $n = 2k + 1$,

$$u_{2k+1}\left(\frac{\pi}{2}, 1\right) = \frac{(-1)^k}{2k+1} e^{-\sqrt{2k+1}} \frac{e^{2k+1} - e^{-(2k+1)}}{2}, \quad (2.23)$$

whose magnitude diverges as $k \rightarrow \infty$. Thus no constant C can satisfy

$$\|u_n\|_{L^\infty(\mathbb{R}_+^2)} \leq C(\|u_n\|_{L^\infty(\partial\mathbb{R}_+^2)} + \|\partial_y u_n\|_{L^\infty(\partial\mathbb{R}_+^2)}). \quad (2.24)$$

A maximum principle fails, and the Cauchy problem is unstable. Figure 2.4 plots the same family along the line $x = \pi/2$. By contrast, the Dirichlet problem for $\Delta u = 0$ on a bounded domain with continuous boundary values is well-posed. The same operator can therefore be well-posed or ill-posed according to the data. Which data are admissible is fixed by the *type* of the equation, developed next.

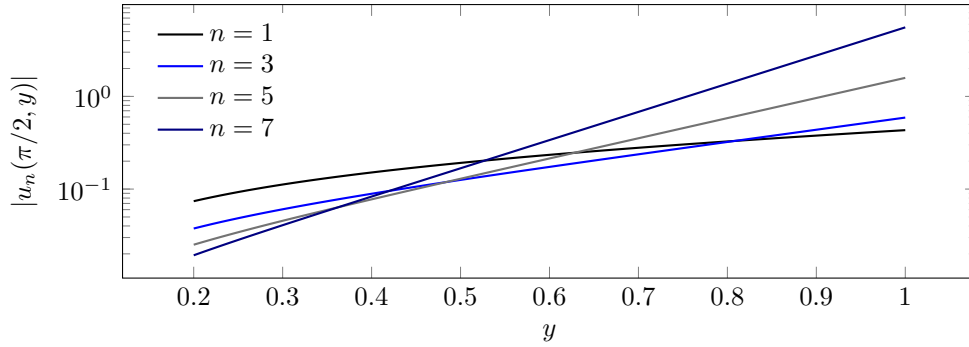


Figure 2.4: Sections of the harmonic functions (2.22) at $x = \pi/2$, on a logarithmic scale. Near $y = 0$ the curves follow the Cauchy data, which decrease with n ; by $y = 1$ the order has reversed and the values grow with n .

A simpler failure of continuous dependence is differentiation itself. The functions $f_n(x) = \sin(nx)/n$ satisfy $\max |f_n| = 1/n \rightarrow 0$ and $\max |f'_n| = \max |\cos(nx)| = 1$. No constant C makes $\max |f'| \leq C \max |f|$ valid for all smooth f , so the map $f \mapsto f'$ is unbounded in the maximum norm: arbitrarily small data can have a derivative of size one. The Cauchy problem for Laplace's equation exhibits the same phenomenon through the factor $\sinh(ny)$. Its discrete counterpart is the amplification of a perturbation of size δ by δ/h in a difference quotient and by δ/h^2 in a second difference, which is the origin of the h^{-2} growth of the condition number in Section 5.14.

2.5 Types of second-order equations

Linear second-order equations are classified by the quadratic form of their highest-order terms, exactly as conic sections are classified by $ax^2 + 2hxy + by^2$ [30]. The type is an invariant of one operator; it decides the correct data and the qualitative behaviour of solutions.

2.5.1 Principal symbol and the two-variable discriminant

In two independent variables the general linear second-order equation is

$$a_{11}u_{xx} + 2a_{12}u_{xy} + a_{22}u_{yy} + a_1u_x + a_2u_y + a_3u = 0, \quad (2.25)$$

with real coefficients, not all of a_{11}, a_{12}, a_{22} equal to zero. (The lower-order coefficients a_1, a_2, a_3 may depend on (x, y) ; they do not enter the type.) Replacing $\partial_x \mapsto \xi$ and $\partial_y \mapsto \eta$ in the highest-order part produces the *principal symbol*, the homogeneous quadratic

$$Q(\xi, \eta) = a_{11}\xi^2 + 2a_{12}\xi\eta + a_{22}\eta^2 = \begin{pmatrix} \xi & \eta \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix}. \quad (2.26)$$

Write A for the symmetric matrix of second-order coefficients and $\delta = \det A = a_{11}a_{22} - a_{12}^2$ for its discriminant. The type of (2.25) is the signature of Q , equivalently the sign of δ .

To see the reduction explicitly, assume first $a_{11} \neq 0$. Expand the square

$$\left(\partial_x + \frac{a_{12}}{a_{11}}\partial_y\right)^2 u = u_{xx} + 2\frac{a_{12}}{a_{11}}u_{xy} + \frac{a_{12}^2}{a_{11}^2}u_{yy}. \quad (2.27)$$

Multiplying by a_{11} and rearranging gives the principal part as

$$\begin{aligned} a_{11}u_{xx} + 2a_{12}u_{xy} + a_{22}u_{yy} &= a_{11}\left(\partial_x + \frac{a_{12}}{a_{11}}\partial_y\right)^2 u + \left(a_{22} - \frac{a_{12}^2}{a_{11}}\right)u_{yy} \\ &= a_{11}X^2u + \frac{\delta}{a_{11}}Y^2u, \end{aligned} \quad (2.28)$$

where $X = \partial_x + (a_{12}/a_{11})\partial_y$ and $Y = \partial_y$. The fields X and Y commute and are linearly independent, so there are coordinates (ξ, η) with $X = \partial_\xi$ and $Y = \partial_\eta$. The choice

$$\xi = x, \quad \eta = y - \frac{a_{12}}{a_{11}}x \quad (2.29)$$

works: the chain rule gives $u_x = u_\xi - (a_{12}/a_{11})u_\eta$ and $u_y = u_\eta$, hence $Xu = u_\xi$ and $Yu = u_\eta$. In these coordinates (2.28) is

$$a_{11}u_{\xi\xi} + \frac{\delta}{a_{11}}u_{\eta\eta}. \quad (2.30)$$

A further dilation of η equalizes the two coefficients. Set $\tilde{\eta} = \lambda\eta$ with $\lambda > 0$, so $u_{\eta\eta} = \lambda^2 u_{\tilde{\eta}\tilde{\eta}}$. The principal part becomes $a_{11}u_{\xi\xi} + (\delta\lambda^2/a_{11})u_{\tilde{\eta}\tilde{\eta}}$. Three cases remain [30].

- If $\delta > 0$, choose $\lambda = |a_{11}|/\sqrt{\delta}$. Then $\delta\lambda^2/a_{11} = a_{11}$, and the principal part is $a_{11}(u_{\xi\xi} + u_{\tilde{\eta}\tilde{\eta}})$. Dividing the PDE by a_{11} (or multiplying by -1 if $a_{11} < 0$) produces $u_{\xi\xi} + u_{\tilde{\eta}\tilde{\eta}} + \dots = 0$, the Laplace operator plus lower-order terms: the equation is *elliptic*.
- If $\delta < 0$, choose $\lambda = |a_{11}|/\sqrt{|\delta|}$. Then $\delta\lambda^2/a_{11} = -a_{11}$, and the principal part is $a_{11}(u_{\xi\xi} - u_{\tilde{\eta}\tilde{\eta}})$. The same normalization produces $u_{\xi\xi} - u_{\tilde{\eta}\tilde{\eta}} + \dots = 0$, the one-dimensional wave operator: the equation is *hyperbolic*.
- If $\delta = 0$, the principal part collapses to $a_{11}u_{\xi\xi}$. The equation is *parabolic* provided a first-order term in η survives (after renaming η as time one obtains the heat operator $u_{\xi\xi} - u_\eta + \dots = 0$). If the first-order part vanishes as well, the equation is degenerate and is no longer of the three standard types.

If $a_{11} = 0$ but $a_{22} \neq 0$, interchange x and y and repeat. If $a_{11} = a_{22} = 0$, then necessarily $a_{12} \neq 0$ and $\delta = -a_{12}^2 < 0$: the principal part is $2a_{12}u_{xy}$, already hyperbolic. The rotation $\xi = (x + y)/\sqrt{2}$, $\eta = (x - y)/\sqrt{2}$ gives $u_x = (u_\xi + u_\eta)/\sqrt{2}$, $u_y = (u_\xi - u_\eta)/\sqrt{2}$, and therefore $u_{xy} = \frac{1}{2}(u_{\xi\xi} - u_{\eta\eta})$, which is the same wave operator as above.

The sign of δ cannot be changed by a real invertible linear substitution $(x, y)^\top = P(\xi, \eta)^\top$. Under that change, $\nabla_{(x, y)} = P^{-T} \nabla_{(\xi, \eta)}$, so the principal part becomes the quadratic form with matrix $P^{-1} A P^{-T}$. Then

$$\det(P^{-1} A P^{-T}) = \frac{\det A}{(\det P)^2}, \quad (2.31)$$

and $(\det P)^2 > 0$, hence $\text{sgn } \delta$ is invariant. Mixed derivatives can always be removed by the calculation above; they never create a fourth type. The three models (2.9) are the constant-coefficient normal forms of the three classes.

2.5.2 Higher dimensions

In n variables the general linear second-order equation is [30]

$$\sum_{i, j=1}^n a_{ij} u_{x_i x_j} + \sum_{i=1}^n a_i u_{x_i} + a_0 u = 0, \quad a_{ij} \in \mathbb{R}. \quad (2.32)$$

As in two variables, the type is read from the principal part alone; the lower-order coefficients a_i and a_0 may be set to zero for the calculation that follows. Collecting the second-order coefficients as they stand produces a real $n \times n$ matrix that need not be symmetric,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \quad (2.33)$$

so that the principal part is $\sum_{i, j} a_{ij} u_{x_i x_j} = \nabla_x^\top A \nabla_x u$. For a C^2 unknown one has $u_{x_i x_j} = u_{x_j x_i}$, hence the antisymmetric part of A drops out:

$$\sum_{i, j} a_{ij} u_{x_i x_j} = \sum_{i, j} \frac{a_{ij} + a_{ji}}{2} u_{x_i x_j} = \nabla_x^\top \left(\frac{A + A^\top}{2} \right) \nabla_x u. \quad (2.34)$$

Replacing A by its symmetric part $(A + A^\top)/2$ does not change the PDE, and one may therefore assume from here on that $a_{ij} = a_{ji}$,

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{12} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{pmatrix}. \quad (2.35)$$

(When $n = 2$ this is the matrix already written in (2.26), with the mixed term recorded as $2a_{12}u_{xy}$.) A linear change $\xi = Bx$ with $B = (b_{kj}) \in M_{n \times n}$ invertible,

$$\begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} = \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \xi_k = \sum_{j=1}^n b_{kj} x_j, \quad (2.36)$$

transforms derivatives by the chain rule

$$\partial_{x_i} = \sum_{k=1}^n \frac{\partial \xi_k}{\partial x_i} \partial_{\xi_k} = \sum_{k=1}^n b_{ki} \partial_{\xi_k}, \quad (2.37)$$

or, in matrix form, $\nabla_x = B^\top \nabla_\xi$. Second derivatives therefore factor as

$$\partial_{x_i} \partial_{x_j} u = \left(\sum_{k=1}^n b_{ki} \partial_{\xi_k} \right) \left(\sum_{\ell=1}^n b_{\ell j} \partial_{\xi_\ell} \right) u. \quad (2.38)$$

Substituting into the principal part gives

$$\sum_{i,j} a_{ij} u_{x_i x_j} = \sum_{i,j} a_{ij} \sum_{k,\ell} b_{ki} b_{\ell j} u_{\xi_k \xi_\ell} = \nabla_\xi^\top (BAB^\top) \nabla_\xi u, \quad (2.39)$$

where the new coefficient matrix has entries

$$(BAB^\top)_{k\ell} = \sum_{i,j=1}^n b_{ki} a_{ij} b_{\ell j}. \quad (2.40)$$

By the spectral theorem there is an orthogonal B (one may take $\det B = 1$) that diagonalizes this form:

$$BAB^\top = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{pmatrix}. \quad (2.41)$$

Then (2.39) collapses to $\sum_k d_k u_{\xi_k \xi_k}$. Equivalently $A = QDQ^\top$ and $\xi = Q^\top x$ produce $\nabla_x^\top A \nabla_x = \nabla_\xi^\top D \nabla_\xi$. The type is the signature of (d_k) , which is invariant because a further invertible change conjugates A to a matrix of the same inertia (Sylvester; in two variables this is (2.31)):

- *elliptic*: every $d_k \neq 0$ and all of one sign (Laplace, Poisson $-\Delta u = f$);
- *hyperbolic*: both signs occur, typically one time-like direction against $n - 1$ space-like ones (wave, Klein–Gordon);
- *parabolic*: exactly one eigenvalue vanishes and the rest share a sign (heat, Fokker–Planck);
- *ultrahyperbolic*: more than one eigenvalue of each sign, which this report does not use.

If the coefficients a_{ij} depend on x , the type is a pointwise attribute of the principal symbol and may change across the domain (Tricomi’s equation $u_{xx} + xu_{yy} = 0$ is elliptic for $x > 0$ and hyperbolic for $x < 0$). The equation remains a single second-order operator; only the signature of $A(x)$ varies.

2.5.3 Characteristics and propagation

The same quadratic form Q determines the *characteristic* curves. A covector (ξ, η) is characteristic when $Q(\xi, \eta) = 0$. After the reduction (2.30) this is $a_{11}\alpha^2 + (\delta/a_{11})\beta^2 = 0$ in the dual coordinates (α, β) to (ξ, η) .

- If $\delta > 0$, both coefficients have the same sign, so the only real solution is $\alpha = \beta = 0$. There are no real characteristic directions.
- If $\delta < 0$, one obtains two distinct real ratios $\beta/\alpha = \pm |a_{11}|/\sqrt{|\delta|}$, hence two families of characteristic curves.
- If $\delta = 0$, the equation reduces to $\alpha^2 = 0$: a repeated real characteristic direction (the η -axis in (2.29)).

Equivalently, in the original variables the slopes $p = dy/dx$ of characteristic curves solve $Q(p, -1) = 0$, i.e. $a_{11}p^2 - 2a_{12}p + a_{22} = 0$, whose discriminant is -4δ . The geometric consequences are as follows.

- an elliptic equation has no real characteristics. Data cannot propagate along a distinguished family of curves; the solution at an interior point depends on the whole boundary. There is no preferred time. The natural problem is a boundary-value problem.
- a hyperbolic equation has two distinct real characteristic families (the lines $x \pm ct = \text{const}$ for the one-dimensional wave equation). Signals travel at finite speed, domains of dependence are compact, and energy is typically conserved. The natural problem is an initial-value problem with Cauchy data (ϕ, ψ) .
- a parabolic equation has a repeated real characteristic (the lines $t = \text{const}$ for the heat equation). Information spreads instantaneously over the whole space, but with exponential damping of high frequencies. The natural problem is an initial-boundary-value problem with a single initial datum ϕ .

Imposing Cauchy data on an elliptic equation, or dropping u_t from a wave equation, is precisely what produces the ill-posed examples of Section 2.4.

2.5.4 One operator, three signatures

The classification is a change of coordinates on a single second-order linear operator. After the reduction, the only invariant is the sign of δ (or the inertia of A in n variables). Laplace, heat, and wave are the constant-coefficient normal forms of that one object, not three separate theories. Characteristics then decide which data are natural, which is why the same Laplacian is well-posed as a Dirichlet problem and ill-posed as a Cauchy problem (Section 2.4). Chapter 3 writes each normal form as a convolution against a kernel. Later chapters discretize that kernel and keep this classification.

2.6 Analyticity and the Laplace equation

In two dimensions a complex function is a real map of the plane. Write $z = x + iy$ and $f = u + iv$, and regard f as $F(x, y) = (u(x, y), v(x, y))$. The complex derivative at z_0 is the limit

$$f'(z_0) = \lim_{h \rightarrow 0} \frac{f(z_0 + h) - f(z_0)}{h}, \quad (2.42)$$

and the limit must be the same along every path $h \rightarrow 0$ [2]. Two paths already force a relation between the real partials. Along the real axis, $h = t \in \mathbb{R}$, the difference quotient is the real partial of f :

$$f'(z_0) = f_x = u_x + iv_x. \quad (2.43)$$

Along the imaginary axis, $h = it$ with $t \in \mathbb{R}$, the same quotient is $(1/i)f_y$. Since $1/i = -i$,

$$\frac{1}{i}f_y = -i(u_y + iv_y) = v_y - iu_y. \quad (2.44)$$

Equating (2.43) and (2.44) gives $u_x + iv_x = v_y - iu_y$. Real and imaginary parts match separately, which is the Cauchy–Riemann system

$$u_x = v_y, \quad v_x = -u_y \quad (2.45)$$

(the second identity is the same as $u_y = -v_x$). If either path produced a different value, (2.42) would not exist. As a first-order system for the vector (u, v) ,

$$\begin{pmatrix} \partial_x & -\partial_y \\ \partial_y & \partial_x \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \quad (2.46)$$

The first row is $u_x - v_y = 0$ and the second is $u_y + v_x = 0$. The Jacobian of F is then multiplication by the complex number f' :

$$J_F = \begin{pmatrix} u_x & u_y \\ v_x & v_y \end{pmatrix} = \begin{pmatrix} u_x & u_y \\ -u_y & u_x \end{pmatrix}, \quad (2.47)$$

so $\det J_F = u_x^2 + u_y^2 = |f'|^2$.

The Wirtinger operators are the change of coordinates that packages the two paths into one complex equation. Solve $z = x + iy$ and $\bar{z} = x - iy$ for the real variables, $x = (z + \bar{z})/2$ and $y = (z - \bar{z})/(2i)$. The chain rule, treating z and $\bar{z}$ as the independent coordinates, is

$$\frac{\partial}{\partial z} = \frac{\partial x}{\partial z} \partial_x + \frac{\partial y}{\partial z} \partial_y, \quad \frac{\partial x}{\partial z} = \frac{1}{2}, \quad \frac{\partial y}{\partial z} = \frac{1}{2i}. \quad (2.48)$$

Since $1/i = -i$ one has $1/(2i) = -i/2$, and likewise $\partial x/\partial \bar{z} = 1/2$, $\partial y/\partial \bar{z} = -1/(2i)$. Therefore

$$\frac{\partial}{\partial z} = \frac{1}{2} \partial_x + \frac{1}{2i} \partial_y = \frac{1}{2} \left(\partial_x + \frac{1}{i} \partial_y \right) = \frac{1}{2} (\partial_x - i \partial_y), \quad (2.49)$$

$$\frac{\partial}{\partial \bar{z}} = \frac{1}{2} \partial_x - \frac{1}{2i} \partial_y = \frac{1}{2} \left(\partial_x - \frac{1}{i} \partial_y \right) = \frac{1}{2} (\partial_x + i \partial_y). \quad (2.50)$$

These are not new derivatives of f . They are linear combinations of ∂_x and ∂_y . Applied to f , the combination $\partial f/\partial z$ is the average of the two difference quotients (2.43) and (2.44), and $\partial f/\partial \bar{z}$ is half their difference:

$$\frac{\partial f}{\partial z} = \frac{1}{2} \left(f_x + \frac{1}{i} f_y \right), \quad \frac{\partial f}{\partial \bar{z}} = \frac{1}{2} \left(f_x - \frac{1}{i} f_y \right) = \frac{1}{2} (u_x - v_y + i(v_x + u_y)). \quad (2.51)$$

The two quotients agree if and only if the difference vanishes, so holomorphy is exactly $\partial f/\partial \bar{z} = 0$, which unpacks into (2.45). When that holds, $f_x + (1/i)f_y = (u_x + iv_x) + (v_y - iu_y) = 2u_x + 2iv_x$, and the average collapses to

$$f' = \frac{\partial f}{\partial z} = \frac{1}{2} (2u_x + 2iv_x) = u_x + iv_x, \quad (2.52)$$

with the rewriting $f' = u_x - iu_y$ from $v_x = -u_y$. The factor 2 sits inside the average and cancels; it is not a separate formula for f' .

Laplace's equation is what remains after one more differentiation. Differentiate $u_x = v_y$ in x and $u_y = -v_x$ in y , and add:

$$\Delta u = u_{xx} + u_{yy} = v_{yx} - v_{xy} = 0, \quad (2.53)$$

the mixed derivatives cancelling for $v \in C^2$. The same argument with the roles reversed gives $\Delta v = 0$. Thus each component of a holomorphic function is harmonic: in two dimensions, Laplace's equation is the real form of holomorphy. The converse, a harmonic conjugate on a disk, the identification of the circle mean with Cauchy's integral formula, and the conformal maps that preserve this equation are proved in Chapter 4. The operator Δ itself is already the stationary heat equation and the elliptic normal form of Section 2.5.

Chapter 3

Analytic solution formulas

The three normal forms of Chapter 2 have explicit solutions when the domain is the whole space, a half-space, a ball, or an interval [30, 21]. In each case the solution is the data integrated against a kernel fixed by the operator. The heat kernel, the Newtonian potential and its boundary correction, and the forward fundamental solution of the wave equation are the continuous forms of the factors e^{tA} , A^{-1} , and e^{-itH} in Table 1.1. This chapter derives those kernels. Chapter 4 reads the two-dimensional ones as contour integrals. Chapter 5 replaces the integral by a matrix.

3.1 The heat kernel

3.1.1 A self-similar solution

The Cauchy problem for the heat equation on the line is

$$u_t = u_{xx}, \quad x \in \mathbb{R}, \quad t > 0, \quad u(x, 0) = \phi(x). \quad (3.1)$$

The equation is invariant under the parabolic scaling $(x, t) \mapsto (\lambda x, \lambda^2 t)$. If u solves $u_t = u_{xx}$ for $t > 0$, so does

$$u_\lambda(x, t) = \lambda u(\lambda x, \lambda^2 t), \quad \lambda > 0. \quad (3.2)$$

The factor λ compensates the change of scale in the integral. Substituting $y = \lambda x$, so that $dx = dy/\lambda$, gives

$$\int_{\mathbb{R}} u_\lambda(x, t) dx = \int_{\mathbb{R}} u(y, \lambda^2 t) dy. \quad (3.3)$$

Integrating the PDE, $\frac{d}{dt} \int_{\mathbb{R}} u dx = \int_{\mathbb{R}} u_{xx} dx = 0$ whenever u_x vanishes at infinity, so the mass does not depend on time and the right-hand side of (3.3) equals $\int_{\mathbb{R}} u(y, t) dy$. A solution fixed by every such rescaling satisfies $u = u_\lambda$. Setting $\lambda = 1/\sqrt{t}$ leaves a profile of the single variable $y = x/\sqrt{t}$. Seek

$$u(x, t) = \frac{c}{\sqrt{t}} V(y), \quad y = \frac{x}{\sqrt{t}}, \quad (3.4)$$

with a constant c still free. The chain rule uses

$$\frac{\partial}{\partial t} \left(\frac{1}{\sqrt{t}} \right) = -\frac{1}{2t^{3/2}}, \quad \frac{\partial y}{\partial t} = x \cdot \left(-\frac{1}{2t^{3/2}} \right) = -\frac{y}{2t}, \quad \frac{\partial y}{\partial x} = \frac{1}{\sqrt{t}}. \quad (3.5)$$

Differentiating the ansatz,

$$\begin{aligned} u_t &= c \left(-\frac{1}{2t^{3/2}} \right) V + \frac{c}{\sqrt{t}} V'(y) \left(-\frac{y}{2t} \right) = -\frac{c}{2t^{3/2}} (V + yV'), \\ u_x &= \frac{c}{\sqrt{t}} V'(y) \cdot \frac{1}{\sqrt{t}} = \frac{c}{t} V', \\ u_{xx} &= \frac{c}{t} V'' \cdot \frac{1}{\sqrt{t}} = \frac{c}{t^{3/2}} V''. \end{aligned} \quad (3.6)$$

The heat equation $u_t = u_{xx}$ is therefore

$$-\frac{c}{2t^{3/2}} (V + yV') = \frac{c}{t^{3/2}} V''. \quad (3.7)$$

Dividing through by $c/t^{3/2}$, for $t > 0$ and $c \neq 0$, leaves the ordinary equation

$$V'' + \frac{y}{2} V' + \frac{1}{2} V = 0. \quad (3.8)$$

The left-hand side is an exact derivative,

$$\left(V' + \frac{y}{2} V \right)' = V'' + \frac{1}{2} V + \frac{y}{2} V' = 0, \quad (3.9)$$

so $V' + (y/2)V$ is constant. A profile that decays at infinity, together with its derivative, forces the constant to vanish. Separating variables,

$$\frac{dV}{V} = -\frac{y}{2} dy, \quad \log V = -\frac{y^2}{4}, \quad V(y) = \exp\left(-\frac{y^2}{4}\right). \quad (3.10)$$

The multiplicative constant in V has been absorbed into c . Substituting $y = x/\sqrt{t}$ gives

$$u(x, t) = \frac{c}{\sqrt{t}} \exp\left(-\frac{x^2}{4t}\right). \quad (3.11)$$

The constant is fixed by conservation of mass. The change of variables $z = x/(2\sqrt{t})$ has $dx = 2\sqrt{t} dz$, and

$$\frac{1}{\sqrt{t}} \exp\left(-\frac{x^2}{4t}\right) dx = \frac{1}{\sqrt{t}} \cdot e^{-z^2} \cdot 2\sqrt{t} dz = 2 e^{-z^2} dz. \quad (3.12)$$

Since $\int_{-\infty}^{\infty} e^{-z^2} dz = \sqrt{\pi}$,

$$\int_{\mathbb{R}} \frac{1}{\sqrt{t}} \exp\left(-\frac{x^2}{4t}\right) dx = 2\sqrt{\pi} = \sqrt{4\pi}. \quad (3.13)$$

The choice $c = 1/\sqrt{4\pi}$ makes the integral equal to one, because $c\sqrt{4\pi} = 1$. The resulting function is the one-dimensional heat kernel

$$\Gamma(x, t) = \frac{1}{\sqrt{4\pi t}} \exp\left(-\frac{x^2}{4t}\right), \quad t > 0. \quad (3.14)$$

It solves (3.1) for $t > 0$, and its integral over the line is 1 at every positive time.

In n space dimensions the same Gaussian factors. The Laplacian is a sum of pure second derivatives, and

$$\Gamma(x, t) = \frac{1}{(4\pi t)^{n/2}} \exp\left(-\frac{|x|^2}{4t}\right) = \prod_{i=1}^n \Gamma_1(x_i, t), \quad (3.15)$$

where the prefactor is the product of the one-dimensional prefactors,

$$\frac{1}{(4\pi t)^{n/2}} = \left(\frac{1}{\sqrt{4\pi t}} \right)^n. \quad (3.16)$$

Each factor solves the heat equation in its own coordinate and integrates to one, so the product solves $u_t = \Delta u$ and integrates to one over $\mathbb{R}^n$. If the equation is $u_t = k\Delta u$ with diffusivity $k > 0$, the same derivation with time scaled by k replaces (3.15) by

$$\Gamma_k(x, t) = \frac{1}{(4\pi kt)^{n/2}} \exp\left(-\frac{|x|^2}{4kt}\right). \quad (3.17)$$

The product form is a shortcut that hides the mechanism. The self-similar derivation also works directly in $\mathbb{R}^n$. Mass conservation and the scaling $(x, t) \mapsto (\lambda x, \lambda^2 t)$ now require the prefactor λ^n in place of λ , so the ansatz is $u(x, t) = t^{-n/2} V(y)$ with $y = |x|/\sqrt{t}$. For a radial function $\Delta = \partial_r^2 + \frac{n-1}{r}\partial_r$, and with $r = y\sqrt{t}$ the chain rule gives

$$u_t = -\frac{1}{2} t^{-\frac{n}{2}-1} (nV + yV'), \quad \Delta u = t^{-\frac{n}{2}-1} \left(V'' + \frac{n-1}{y} V' \right). \quad (3.18)$$

Equating them leaves $V'' + \frac{n-1}{y} V' + \frac{1}{2} (nV + yV') = 0$. Multiplication by y^{n-1} turns the left side into an exact derivative,

$$\left(y^{n-1} V' + \frac{1}{2} y^n V \right)' = 0. \quad (3.19)$$

The quantity $y^{n-1} (V' + \frac{y}{2} V)$ is constant, and it tends to 0 as $y \rightarrow 0$ for a smooth profile, so the constant is zero and $V' = -\frac{y}{2} V$. Hence $V = C e^{-y^2/4}$. Since $\int_{\mathbb{R}^n} e^{-|x|^2/4t} dx = (4\pi t)^{n/2}$, unit mass fixes $C = (4\pi)^{-n/2}$, which is (3.15). For $n = 1$ the profile equation is (3.8). The factor $(n-1)/y$ is the only trace of the dimension, as the term $(n-1)/r$ will be in the radial wave equation (3.112). Figure 3.1 shows the one-dimensional kernel at three times: the mass stays equal to one while the peak falls and the width grows like $\sqrt{t}$.

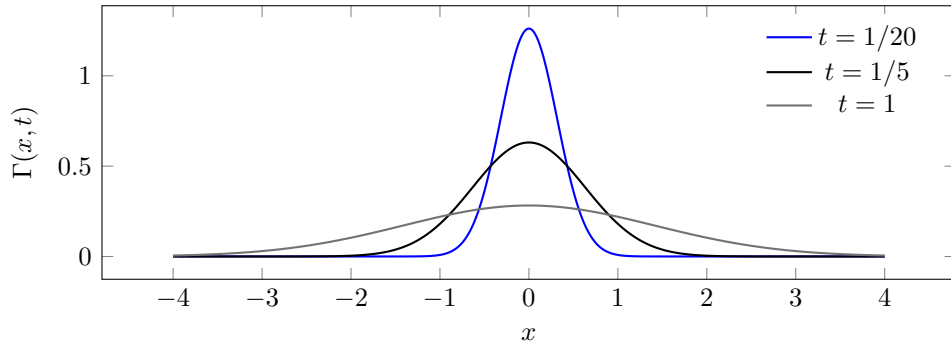


Figure 3.1: The heat kernel (3.14) at three times. Each graph has integral one. As t decreases the peak rises and the mass concentrates at the origin, which is the approach to δ .

3.1.2 Convolution against the initial datum

The kernel produces the solution of (3.1) for general initial data. Translation in x preserves the equation, so $\Gamma(x - y, t)$ solves the heat equation for each fixed y . An integral of solutions against

a density ϕ is again a solution:

$$u(x, t) = \int_{\mathbb{R}^n} \Gamma(x - y, t) \phi(y) dy = (\Gamma(\cdot, t) * \phi)(x). \quad (3.20)$$

This is the semigroup $e^{t\Delta}$ applied to ϕ . It is the same pattern as Duhamel's formula (2.3), with the scalar factor $e^{\lambda(x-t)}$ replaced by the kernel of $e^{t\Delta}$.

If ϕ is bounded and continuous, $u(\cdot, t)$ converges to ϕ uniformly on compact sets as $t \rightarrow 0$. Because $\int \Gamma(\cdot, t) = 1$,

$$u(x, t) - \phi(x) = \int_{\mathbb{R}^n} \frac{1}{(4\pi t)^{n/2}} \exp\left(-\frac{|x-y|^2}{4t}\right) (\phi(y) - \phi(x)) dy. \quad (3.21)$$

Fix $\varepsilon > 0$ and choose $\delta > 0$ so that $|\phi(y) - \phi(x)| < \varepsilon$ whenever $|y - x| < \delta$, uniformly for x in the compact set under consideration. Split the integral at that distance:

$$\begin{aligned} |u(x, t) - \phi(x)| &\leq \int_{|x-y| \leq \delta} \frac{1}{(4\pi t)^{n/2}} \exp\left(-\frac{|x-y|^2}{4t}\right) |\phi(y) - \phi(x)| dy \\ &\quad + \int_{|x-y| > \delta} \frac{1}{(4\pi t)^{n/2}} \exp\left(-\frac{|x-y|^2}{4t}\right) |\phi(y) - \phi(x)| dy. \end{aligned} \quad (3.22)$$

The first integrand is at most ε times the kernel, and the kernel integrates to at most 1, so the near integral is at most ε . In the far integral, $|\phi(y) - \phi(x)| \leq 2\|\phi\|_\infty$. The change of variables $z = (x - y)/(2\sqrt{t})$ has $dy = (2\sqrt{t})^n dz$, and

$$\frac{(2\sqrt{t})^n}{(4\pi t)^{n/2}} = \frac{1}{\pi^{n/2}}, \quad (3.23)$$

so the threshold $|x - y| = \delta$ becomes $|z| = \delta/(2\sqrt{t})$ and the far integral is at most

$$2\|\phi\|_\infty \int_{|z| > \delta/(2\sqrt{t})} \frac{1}{\pi^{n/2}} e^{-|z|^2} dz. \quad (3.24)$$

As $t \rightarrow 0$ the region of integration escapes to infinity and the integral of the fixed Gaussian tends to zero. For t small enough the tail is less than ε , and $|u(x, t) - \phi(x)| < 2\varepsilon$. Thus $\Gamma(\cdot, t)$ converges to the Dirac mass δ at the origin, in the sense that convolution against it reproduces continuous data.

If instead $\phi \in L^1(\mathbb{R}^n)$, the same splitting works in L^1 . Continuity of translation in L^1 makes the near region small, and the Gaussian tail makes the far region small, so $\|u(\cdot, t) - \phi\|_{L^1} \rightarrow 0$. The pointwise size of the kernel gives the bound

$$\|u(\cdot, t)\|_\infty \leq \frac{1}{(4\pi t)^{n/2}} \|\phi\|_1. \quad (3.25)$$

Both bounds are cases of Young's convolution inequality, $\|f * g\|_{L^p} \leq \|f\|_{L^q} \|g\|_{L^r}$ for $\frac{1}{p} + 1 = \frac{1}{q} + \frac{1}{r}$. Since $\Gamma \geq 0$ and $\int \Gamma(\cdot, t) = 1$ one has $\|\Gamma(\cdot, t)\|_{L^1} = 1$, and the choice $r = 1$, $q = p$ gives

$$\|u(\cdot, t)\|_{L^p} \leq \|\phi\|_{L^p}, \quad 1 \leq p \leq \infty. \quad (3.26)$$

The heat flow contracts every L^p norm. The case $p = \infty$ is the maximum principle on the whole space and $p = 2$ is the energy inequality (3.34); the earlier bound is Young with $g = \Gamma \in L^\infty$ and $f = \phi \in L^1$.

The same scaling gives smoothing. From $\Gamma(x, t) = t^{-n/2} \Gamma(x/\sqrt{t}, 1)$ and the fact that every derivative of $\Gamma(\cdot, 1)$ is a polynomial times a Gaussian,

$$|\partial_x^\alpha \Gamma(x, t)| \leq \frac{C_\alpha}{(\sqrt{t} + |x|)^{n+|\alpha|}}, \quad \|\partial_x^\alpha \Gamma(\cdot, t)\|_{L^1} = C'_\alpha t^{-|\alpha|/2}. \quad (3.27)$$

The right side is integrable in y for $t > 0$, so derivatives of every order may be taken under the integral in (3.20). The function $u(\cdot, t)$ is C^∞ for $t > 0$ even when ϕ is only bounded or integrable, and this justifies the statement that an integral of solutions against a density is again a solution. The constant $t^{-|\alpha|/2}$ blows up as $t \rightarrow 0$ and records that derivatives of rough data are large. Each time derivative costs Δ , hence two powers of $\sqrt{t}$.

A forcing term is Duhamel's formula in space and time. The solution of $u_t = \Delta u + f$ with initial value ϕ is

$$u(x, t) = \int_{\mathbb{R}^n} \Gamma(x - y, t) \phi(y) dy + \int_0^t \int_{\mathbb{R}^n} \Gamma(x - y, t - s) f(y, s) dy ds. \quad (3.28)$$

The inner integral at each frozen time s is (3.20) applied to the slice $f(\cdot, s)$, and the outer integral superposes those contributions after they have diffused for time $t - s$. On a bounded domain the same formula holds with Γ replaced by the Green's function (3.71), which subtracts a corrector so that the boundary condition is met.

3.1.3 The maximum principle

On a bounded interval the maximum of a solution cannot appear in the interior of the space-time rectangle. Let $Q_T = (0, L) \times (0, T]$ and write $\partial_p Q_T$ for the parabolic boundary, consisting of the bottom $\{t = 0\}$ and the lateral sides $\{x = 0\}$ and $\{x = L\}$.

$$\max_{\overline{Q_T}} u = \max_{\partial_p Q_T} u \quad (3.29)$$

whenever $u_t = u_{xx}$ in Q_T and u is continuous up to the closure. The argument compares u with $v = u + \varepsilon x^2$. Then $v_t - v_{xx} = -2\varepsilon < 0$. At an interior maximum, or at a maximum on the open top $\{t = T\} \times (0, L)$, one would have $v_t \geq 0$ and $v_{xx} \leq 0$, hence $v_t - v_{xx} \geq 0$, which contradicts $-2\varepsilon < 0$. The maximum of v is therefore attained on $\partial_p Q_T$, where it is at most $\max_{\partial_p Q_T} u + \varepsilon L^2$. Since $u = v - \varepsilon x^2$,

$$u \leq \max_{\partial_p Q_T} u + \varepsilon L^2 \quad (3.30)$$

everywhere in $\overline{Q_T}$. The parameter $\varepsilon > 0$ is arbitrary, so (3.29) follows. The same statement for $-u$ is the minimum principle. A solution of the initial-boundary-value problem is therefore controlled by the initial datum and the prescribed lateral boundary values, which is the continuous dependence half of well-posedness for this parabolic problem.

The comparison (3.29) is the weak maximum principle. The strong form adds constancy. If u , u_t , and u_{xx} are continuous on the closed rectangle and u attains its maximum at a point (x_0, t_0) with $0 < x_0 < L$ and $0 < t_0 \leq T$, then u is constant on $[0, L] \times [0, t_0]$. An interior maximum is the value u already takes throughout that earlier cylinder.

Uniqueness is the weak principle applied to a difference. Consider

$$u_t = u_{xx} \quad \text{on } Q_T, \quad u(x, 0) = \phi(x), \quad u(0, t) = g(t), \quad u(L, t) = h(t). \quad (3.31)$$

If u_1 and u_2 both solve (3.31), then $v = u_1 - u_2$ solves the heat equation with zero initial values and zero lateral values. The maximum principle gives $v \leq 0$ and the minimum principle gives $v \geq 0$, so $v \equiv 0$.

The argument is not special to one dimension. Let $\Omega \subset B_R(0) \subset \mathbb{R}^n$ be bounded, $\Omega_T = \Omega \times (0, T]$, and let $\partial_p \Omega_T = (\overline{\Omega} \times \{0\}) \cup (\partial\Omega \times [0, T])$. If $u_t = \Delta u$ in Ω_T and u is continuous on the closure, then $\max_{\overline{\Omega_T}} u = \max_{\partial_p \Omega_T} u$. Set $v = u + \varepsilon|x|^2$. Since $\Delta|x|^2 = 2n$, one has $v_t - \Delta v = -2n\varepsilon < 0$, and an interior (or top) maximum would force $v_t \geq 0 \geq \Delta v$, a contradiction. Hence $u \leq \max_{\partial_p \Omega_T} u + \varepsilon R^2$, and $\varepsilon \rightarrow 0$ gives the claim. The same device proves the weak maximum principle for harmonic functions: if $\Delta u = 0$ in Ω , then $v = u + \varepsilon|x|^2$ has $\Delta v = 2n\varepsilon > 0$, which excludes an interior maximum (at a maximum the Hessian is negative semidefinite, so $\Delta v \leq 0$), and again $\max_{\overline{\Omega}} u = \max_{\partial\Omega} u$. A comparison principle follows by applying these statements to a difference: if $u_t - \Delta u \leq w_t - \Delta w$ in Ω_T and $u \leq w$ on $\partial_p \Omega_T$, then $u \leq w$ throughout. In particular nonnegative data produce a nonnegative solution, so the heat flow preserves positivity. On the whole space $\mathbb{R}^n$ the argument needs a growth bound such as $|u| \leq Ce^{a|x|^2}$ at infinity. Without it uniqueness fails (Tychonoff's example), which is why the bounded solutions of Section 3.1.2 are the right class.

The energy identity gives the same uniqueness with an explicit norm. Let w solve the heat equation with $w(0, t) = w(L, t) = 0$. Multiply $w_t = w_{xx}$ by w and integrate:

$$\frac{1}{2} \frac{d}{dt} \int_0^L w(x, t)^2 dx = \int_0^L w w_{xx} dx = [w w_x]_0^L - \int_0^L (w_x)^2 dx = - \int_0^L (w_x)^2 dx \leq 0. \quad (3.32)$$

The boundary term vanishes because w vanishes at the endpoints. Integrating in time,

$$\int_0^L w(x, t)^2 dx \leq \int_0^L w(x, 0)^2 dx. \quad (3.33)$$

Zero initial data remain zero in L^2 . The same inequality is the continuous form of the decaying factor e^{tA} in Table 1.1: with $A = \partial_{xx}$ and Dirichlet conditions, the semigroup does not increase the L^2 norm.

On $\mathbb{R}^n$ the boundary term is absent because of decay at infinity. Suppose $w_t = \Delta w$ and w together with its first derivatives vanish at infinity fast enough that $\int_{\mathbb{R}^n} \partial_i(w \partial_i w) dx = 0$. Multiplication by w and an integration by parts give

$$\frac{1}{2} \frac{d}{dt} \int_{\mathbb{R}^n} w^2 dx = \int_{\mathbb{R}^n} w \Delta w dx = - \int_{\mathbb{R}^n} |\nabla w|^2 dx \leq 0. \quad (3.34)$$

Every spatial derivative $\partial^\alpha w$ solves the heat equation. Under the same decay for that derivative,

$$\frac{1}{2} \frac{d}{dt} \int_{\mathbb{R}^n} |\partial^\alpha w|^2 dx = - \int_{\mathbb{R}^n} |\nabla \partial^\alpha w|^2 dx \leq 0. \quad (3.35)$$

Each derivative is therefore nonincreasing in L^2 .

A positive solution carries an entropy with the same sign. If $w > 0$ and the decay above still holds, then $\partial_t(w \ln w) = (\ln w + 1)w_t = (\ln w + 1)\Delta w$. The integral of Δw vanishes, and $\nabla(\ln w) = \nabla w / w$, so

$$\int_{\mathbb{R}^n} (\ln w) \Delta w dx = - \int_{\mathbb{R}^n} \nabla(\ln w) \cdot \nabla w dx = - \int_{\mathbb{R}^n} \frac{|\nabla w|^2}{w} dx. \quad (3.36)$$

Therefore

$$\frac{d}{dt} \int_{\mathbb{R}^n} w \ln w dx = - \int_{\mathbb{R}^n} \frac{|\nabla w|^2}{w} dx \leq 0. \quad (3.37)$$

The decay of the entropy is also convex in time. Write $S(t) = -\int w \ln w \, dx$ and $I(t) = \int |\nabla w|^2 / w \, dx$, the Fisher information, so that (3.37) reads $S' = I \geq 0$. In one dimension put $p = (\ln w)_x$. Then $w_x = wp$ and $w_{xx} = w(p^2 + p_x)$, and $q = \ln w$ satisfies $q_t = w_{xx}/w = q_{xx} + q_x^2$, so that $p_t = p_{xx} + 2pp_x$. With $I = \int wp^2 \, dx$,

$$I' = \int w_{xx} p^2 + 2 \int w p p_t = \int wp^4 + 5 \int wp^2 p_x + 2 \int w p p_{xx}. \quad (3.38)$$

Two integrations by parts, using $w_x = wp$ and $(wp)_x = wp^2 + wp_x$, give $\int wp^2 p_x = -\frac{1}{3} \int wp^4$ and $\int w p p_{xx} = -\int (wp)_x p_x = \frac{1}{3} \int wp^4 - \int wp_x^2$. The p^4 terms cancel, since $1 - \frac{5}{3} + \frac{2}{3} = 0$, and

$$S''(t) = I'(t) = -2 \int_{\mathbb{R}} w ((\ln w)_{xx})^2 \, dx \leq 0. \quad (3.39)$$

In n dimensions the same computation gives $-2 \int w |\nabla^2 \ln w|^2$, with the Hessian of $\ln w$. The entropy therefore increases and is concave: the production of entropy slows as the profile approaches a Gaussian. For $w = \Gamma$ one has $S = \frac{1}{2} \ln(4\pi e t)$, $S' = 1/(2t)$, and $S'' = -1/(2t^2)$, in agreement with $(\ln \Gamma)_{xx} = -1/(2t)$.

3.1.4 The half-line

Boundary conditions are imposed by reflection, as for the wave equation in Section 3.4.2. Consider $u_t = u_{xx}$ for $x > 0$, $t > 0$, with $u(x, 0) = \phi(x)$ and $u(0, t) = 0$. The compatibility condition $\phi(0) = 0$ is needed for u to be continuous at the corner. Extend the data oddly, $\phi_o(-x) = -\phi(x)$. The whole-line solution $U = \Gamma(\cdot, t) * \phi_o$ is odd in x , because the kernel is even and ϕ_o is odd, and an odd continuous function vanishes at $x = 0$. Its restriction to $x > 0$ solves the half-line problem. Split the integral at $y = 0$ and substitute $y \rightarrow -y$ on the negative part:

$$u(x, t) = \int_0^\infty [\Gamma(x - y, t) - \Gamma(x + y, t)] \phi(y) \, dy. \quad (3.40)$$

For a Neumann condition $u_x(0, t) = 0$ the extension is even, and

$$u(x, t) = \int_0^\infty [\Gamma(x - y, t) + \Gamma(x + y, t)] \phi(y) \, dy. \quad (3.41)$$

Both boundary conditions can be checked directly: at $x = 0$ the bracket in (3.40) is $\Gamma(-y, t) - \Gamma(y, t) = 0$, and the x -derivative of the bracket in (3.41) is $\Gamma_x(-y, t) + \Gamma_x(y, t) = 0$ because Γ_x is odd. The bracket is the heat kernel with the boundary corrector of (3.71): the corrector $\psi^y(x, t) = \pm \Gamma(x + y, t)$ is the kernel of an image source at $-y$, exactly as the mirror point $\tilde{x}$ in (3.77). On a bounded interval $(0, L)$ the reflections at the two ends produce images at $2kL \pm y$ for all integers k , and the kernel is the sum $\sum_k [\Gamma(x - y + 2kL, t) - \Gamma(x + y + 2kL, t)]$. By the Poisson summation formula this sum equals the sine series (3.129). The image sum is efficient for short times, when only the nearest images matter, and the series for long times. On a grid the odd extension about a Dirichlet node is the discrete sine transform and the even extension about a Neumann node is the discrete cosine transform, as recorded in Section 5.10.

3.2 The Newtonian potential

The stationary problem is Poisson's equation $-\Delta u = f$ on $\mathbb{R}^n$, $n \geq 2$. A radial solution of the homogeneous equation $\Delta u = 0$ for $x \neq 0$ depends only on $r = |x|$. In radial coordinates

$$\Delta u = u'' + \frac{n-1}{r} u' = \frac{1}{r^{n-1}} (r^{n-1} u')', \quad (3.42)$$

so $r^{n-1}u'$ is constant. Call the constant A . Then

$$u'(r) = \frac{A}{r^{n-1}}. \quad (3.43)$$

For $n = 2$ this is $u' = A/r$, and one integration gives $u = A \log r$. For $n \geq 3$,

$$u(r) = A \int \frac{1}{r^{n-1}} dr = \frac{A}{(2-n)r^{n-2}} = -\frac{A}{(n-2)r^{n-2}}. \quad (3.44)$$

Write α_n for the volume of the unit ball in $\mathbb{R}^n$, so that the surface area of the unit sphere is $n\alpha_n$. The constant A is fixed by requiring the flux through every small sphere to equal one. On $|x| = \varepsilon$ the radial derivative is A/ε^{n-1} . The outward unit normal to the domain with the ball removed points toward the origin, so

$$\frac{\partial u}{\partial \nu} = -\frac{A}{\varepsilon^{n-1}}. \quad (3.45)$$

Multiplying by the area $n\alpha_n\varepsilon^{n-1}$ gives the flux $-An\alpha_n$. The value 1 forces $A = -1/(n\alpha_n)$. For $n = 2$ one has $n\alpha_2 = 2\pi$, hence $u = -(1/(2\pi)) \log r$. For $n \geq 3$, (3.44) becomes

$$u(r) = \frac{1}{n(n-2)\alpha_n r^{n-2}}. \quad (3.46)$$

The fundamental solution, or Newtonian potential, is this radial profile:

$$\Phi(x) = \begin{cases} -\frac{1}{2\pi} \log |x|, & n = 2, \\ \frac{1}{n(n-2)\alpha_n |x|^{n-2}}, & n \geq 3. \end{cases} \quad (3.47)$$

Both formulas are instances of one gradient. Differentiating (3.43) with $A = -1/(n\alpha_n)$ and writing the radial unit vector as $x/|x|$ gives, for $x \neq 0$,

$$\nabla \Phi(x) = -\frac{1}{n\alpha_n} \frac{x}{|x|^n}. \quad (3.48)$$

In three dimensions $\alpha_3 = 4\pi/3$, so $n(n-2)\alpha_3 = 4\pi$ and $\Phi(x) = 1/(4\pi|x|)$.

The distributional identity is $-\Delta \Phi = \delta$. Test Φ against a compactly supported smooth function f and excise the ball B_ε about the origin. On the excised domain Φ is smooth and harmonic, so Green's second identity, recorded in (3.60) below, leaves two integrals over the sphere $|x| = \varepsilon$. The outward unit normal to the excised domain points toward the origin, $\nu = -x/|x|$. One integral contains Φ times the normal derivative of f . On that sphere $|\Phi|$ is bounded by a constant times $|\log \varepsilon|$ when $n = 2$, and equals $1/(n(n-2)\alpha_n\varepsilon^{n-2})$ when $n \geq 3$. Multiplied by the area $n\alpha_n\varepsilon^{n-1}$, the integral is of order $\varepsilon|\log \varepsilon|$ or of order ε , and both tend to 0 as $\varepsilon \rightarrow 0$. The other integral is the flux of $\nabla \Phi$. Combined with (3.48),

$$\frac{\partial \Phi}{\partial \nu} = \frac{1}{n\alpha_n \varepsilon^{n-1}} \quad \text{on } |x| = \varepsilon, \quad (3.49)$$

so the flux equals 1 for every $\varepsilon > 0$. The spherical average of the test function therefore tends to its value at the origin. Passing to the limit produces

$$-\Delta \Phi = \delta \quad \text{in } \mathbb{R}^n. \quad (3.50)$$

Consequently the Newtonian potential of a compactly supported forcing,

$$u(x) = \int_{\mathbb{R}^n} \Phi(x-y) f(y) dy = (\Phi * f)(x), \quad (3.51)$$

solves $-\Delta u = f$. Formally, $-\Delta u = (-\Delta \Phi) * f = \delta * f = f$, and the excision above is the justification for $f \in C_c^\infty$. Equation (3.51) is the whole-space form of A^{-1} applied to f , before any boundary condition has been imposed.

A harmonic function satisfies the mean-value property. If $\Delta u = 0$ in a domain containing the closed ball $\overline{B_r(x)}$, then

$$u(x) = \frac{1}{|\partial B_r(x)|} \int_{\partial B_r(x)} u dS = \frac{1}{|B_r(x)|} \int_{B_r(x)} u dy. \quad (3.52)$$

Write the spherical mean as $\varphi(r) = \frac{1}{|\partial B_r(x)|} \int_{\partial B_r(x)} u dS$. The change of variables $y = x + rz$ moves the integral to the unit sphere. Differentiating in r and returning to y gives

$$\varphi'(r) = \frac{1}{|\partial B_r(x)|} \int_{\partial B_r(x)} \frac{\partial u}{\partial \nu} dS = \frac{1}{|\partial B_r(x)|} \int_{B_r(x)} \Delta u dy = 0, \quad (3.53)$$

where the second equality is the divergence theorem. Thus φ is constant, and its limit as $r \rightarrow 0$ is $u(x)$. Integrating the spherical means in r gives the solid-ball formula in (3.52). In two dimensions the spherical mean is the real part of Cauchy's integral formula, proved from the integral in Section 4.8. The strong maximum principle is the same identity: a harmonic function on a bounded connected domain that attains its maximum at an interior point equals that maximum on a ball, and a chain of such balls fills the domain, so the maximum is attained everywhere. The maximum of a harmonic function on a bounded domain is therefore attained on the boundary. If two solutions of $-\Delta u = f$ share the same boundary values, their difference is harmonic and vanishes on $\partial\Omega$, so the maximum and minimum principles force the difference to vanish.

3.2.1 Poisson as the time integral of heat

The two kernels are related by $\Phi(x) = \int_0^\infty \Gamma(x, t) dt$ for $n \geq 3$. Substitute $s = |x|^2/(4t)$, so that $t = |x|^2/(4s)$, $dt = -\frac{|x|^2}{4s^2} ds$, and $(4\pi t)^{-n/2} = \pi^{-n/2} |x|^{-n} s^{n/2}$. Then

$$\int_0^\infty \frac{e^{-|x|^2/4t}}{(4\pi t)^{n/2}} dt = \frac{1}{\pi^{n/2} |x|^n} \cdot \frac{|x|^2}{4} \int_0^\infty s^{\frac{n}{2}-2} e^{-s} ds = \frac{\Gamma_E(\frac{n}{2}-1)}{4\pi^{n/2}} |x|^{2-n}, \quad (3.54)$$

where Γ_E is Euler's gamma function and the integral converges at $s = 0$ only if $n \geq 3$. With $\alpha_n = \pi^{n/2}/\Gamma_E(\frac{n}{2}+1)$ and $\Gamma_E(\frac{n}{2}+1) = \frac{n}{2}(\frac{n}{2}-1)\Gamma_E(\frac{n}{2}-1)$ one finds $n(n-2)\alpha_n = 4\pi^{n/2}/\Gamma_E(\frac{n}{2}-1)$, so the right side is $\frac{1}{n(n-2)\alpha_n |x|^{n-2}} = \Phi(x)$. The identity also follows without computation. Since $\Gamma_t = \Delta \Gamma$,

$$-\Delta \int_0^T \Gamma dt = - \int_0^T \Gamma_t dt = \Gamma(\cdot, 0) - \Gamma(\cdot, T) = \delta - \Gamma(\cdot, T), \quad (3.55)$$

and $\Gamma(\cdot, T) \rightarrow 0$ as $T \rightarrow \infty$, so the limit of the integral solves $-\Delta \Phi = \delta$. For $n = 1, 2$ the integrand decays only like $t^{-n/2}$ and the integral diverges at large t ; the logarithmic Φ of $n = 2$ is obtained after subtracting an infinite constant.

This is the continuous form of $A^{-1} = \int_0^\infty e^{-tA} dt$, which for a symmetric positive definite A is the scalar identity $\lambda^{-1} = \int_0^\infty e^{-t\lambda} dt$ applied to every eigenvalue. The heat semigroup $e^{t\Delta}$ has the kernel Γ , the inverse of $-\Delta$ has the kernel Φ , and the second is the time integral of the first:

solving Poisson's equation is following the heat flow to its steady state. The integral converges when the spectrum is bounded away from zero. On $\mathbb{R}^n$ the spectrum of $-\Delta$ reaches 0 and the integral converges only for $n \geq 3$, because of the decay of Γ . On a bounded domain with Dirichlet data the lowest eigenvalue λ_1 is positive, e^{-tA} decays like $e^{-\lambda_1 t}$, and $G(x, y) = \int_0^\infty G(x, y, t) dt$ in every dimension, with $G(x, y, t)$ the Dirichlet heat kernel of (3.71). The decay rate also controls the truncation: stopping the integral at T leaves an error of order $e^{-\lambda_1 T}$, so the horizon grows with the inverse of the smallest eigenvalue, that is, with the condition number. The same representation of A^{-1} as an integral over a semigroup is what linear combination of Hamiltonian simulation exploits in Section 11.7, and the Cauchy integral for $1/z$ in Section 9.5 writes the same inverse as a contour integral of resolvents.

3.3 Green's function

3.3.1 The boundary correction

Green's formulas rest on two theorems of vector calculus, recorded here because the rest of the chapter uses them repeatedly. Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with piecewise smooth boundary and let ν be the outward unit normal, as in Figure 3.2. For a vector field $F \in C^1(\bar{\Omega}; \mathbb{R}^n)$ the divergence theorem states

$$\int_{\Omega} \nabla \cdot F \, dx = \int_{\partial\Omega} F \cdot \nu \, dS. \quad (3.56)$$

The left side adds up the sources of the field inside Ω and the right side is the net flux that leaves through the boundary. It is the conservation argument behind $u_t = k\Delta u$ and behind the spherical-mean identity of Section 3.2. For $\Omega = (a, b)$ in one dimension, $\partial\Omega = \{a, b\}$ with $\nu = -1$ at a and $\nu = +1$ at b , and (3.56) is the fundamental theorem of calculus $F(b) - F(a) = \int_a^b F' \, dx$.

Stokes' theorem is the companion statement for circulation. Let $S \subset \mathbb{R}^3$ be a smooth oriented surface with unit normal ν and boundary curve ∂S , traversed counterclockwise when seen from the tip of ν . For $F \in C^1(\bar{S}; \mathbb{R}^3)$,

$$\int_S (\nabla \times F) \cdot \nu \, dS = \oint_{\partial S} F \cdot dr. \quad (3.57)$$

Take $S = \Omega$ to be a plane region with $\nu = e_3$ and $F = (P, Q, 0)$. Then $(\nabla \times F) \cdot e_3 = Q_x - P_y$, and (3.57) becomes Green's theorem in the plane,

$$\int_{\Omega} (Q_x - P_y) \, dx \, dy = \oint_{\partial\Omega} P \, dx + Q \, dy. \quad (3.58)$$

On a counterclockwise curve with tangent $\tau = (\tau_1, \tau_2)$ the outward normal is $\nu = (\tau_2, -\tau_1)$, so $\nu \, dS = (dy, -dx)$. With $F = (Q, -P)$ the divergence theorem (3.56) reads $\int_{\Omega} (Q_x - P_y) = \oint Q \, dy + P \, dx$. In two dimensions the divergence theorem and Stokes' theorem are therefore one identity written with a normal or with a tangent. For a function $f = u + iv$ the same identity applies to $\oint f \, dz = \oint (u \, dx - v \, dy) + i \oint (v \, dx + u \, dy)$. The area integrands are $-(v_x + u_y)$ and $u_x - v_y$, and both vanish by the Cauchy–Riemann equations (2.45). This gives Cauchy's theorem for C^1 functions, and Chapter 4 proves it without assuming that f' is continuous.

The two integral identities that move derivatives off a solution and onto a kernel are Green's formulas. For $u, v \in C^2(\bar{\Omega})$ and ν the outward unit normal,

$$\int_{\Omega} v \Delta u \, dx = \int_{\partial\Omega} v \frac{\partial u}{\partial \nu} \, dS - \int_{\Omega} \nabla v \cdot \nabla u \, dx, \quad (3.59)$$

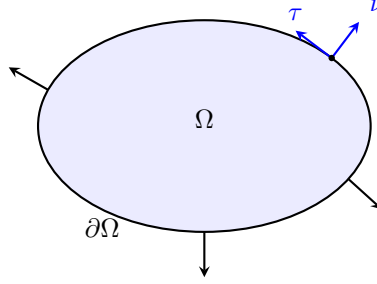


Figure 3.2: A bounded domain Ω with the outward unit normal ν on its boundary. The boundary is traversed counterclockwise, so Ω lies on the left of the tangent τ . In the plane ν is τ turned clockwise by a right angle, which is why (3.56) and (3.58) are one identity.

$$\int_{\Omega} (u \Delta v - v \Delta u) dx = \int_{\partial \Omega} \left(u \frac{\partial v}{\partial \nu} - v \frac{\partial u}{\partial \nu} \right) dS. \quad (3.60)$$

The first is (3.56) applied to $F = v \nabla u$, because $\nabla \cdot (v \nabla u) = \nabla v \cdot \nabla u + v \Delta u$. The second is the first applied twice, with u and v exchanged, and subtracted. On the whole space the boundary term is absent at infinity for the class of f in (3.51), and one recovers convolution against Φ . On a bounded domain the boundary term survives, and the kernel has to be corrected.

The goal is a formula for the solution of the Dirichlet problem

$$-\Delta u = f \quad \text{in } \Omega, \quad u = g \quad \text{on } \partial \Omega, \quad (3.61)$$

in terms of f and g alone. The whole-space kernel Φ does not suffice, and seeing why shows what the correction must do. Fix $x \in \Omega$ and apply (3.60) to u and $v(y) = \Phi(y - x)$ on $\Omega_\varepsilon = \Omega \setminus \overline{B_\varepsilon(x)}$, where $\Delta v = 0$. The left side is $-\int_{\Omega_\varepsilon} \Phi \Delta u dy = \int_{\Omega_\varepsilon} \Phi f dy$. The boundary of Ω_ε consists of $\partial \Omega$ and the sphere $|y - x| = \varepsilon$. On that sphere the outward normal of Ω_ε points toward x , and $\partial_\nu \Phi = 1/(n\alpha_n \varepsilon^{n-1})$ by the flux computation for Φ , while the area is $n\alpha_n \varepsilon^{n-1}$. The integral of $u \partial_\nu \Phi$ over the sphere is therefore the mean of u on it, which tends to $u(x)$. The integral of $\Phi \partial_\nu u$ is of order $\varepsilon |\log \varepsilon|$ for $n = 2$ and of order ε for $n \geq 3$, and tends to 0. In the limit

$$u(x) = \int_{\Omega} \Phi(y - x) f(y) dy - \int_{\partial \Omega} u \frac{\partial \Phi}{\partial \nu} dS + \int_{\partial \Omega} \Phi \frac{\partial u}{\partial \nu} dS. \quad (3.62)$$

On the whole space the two boundary integrals are absent and this is (3.51). On a bounded domain the Dirichlet problem prescribes $u = g$ on $\partial \Omega$ but not the normal derivative $\partial_\nu u$, which is part of what is being solved for. The last integral is unknown, so (3.62) is not yet a solution formula. The cure is to replace $\Phi(\cdot - x)$ by a kernel that vanishes on $\partial \Omega$, so that the term $v \partial_\nu u$ in (3.60) drops out.

Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with smooth boundary. The Green's function $G(x, y)$ for $-\Delta$ with homogeneous Dirichlet conditions is defined, for each fixed $x \in \Omega$, by

$$G(x, y) = \Phi(y - x) - \phi^x(y), \quad (3.63)$$

where the corrector ϕ^x solves

$$\Delta_y \phi^x = 0 \quad \text{in } \Omega, \quad \phi^x(y) = \Phi(y - x) \quad \text{for } y \in \partial \Omega. \quad (3.64)$$

The corrector is harmonic in y , so it does not change $-\Delta_y G = \delta_x$. It cancels $\Phi(\cdot - x)$ on the boundary, so $G(x, \cdot) = 0$ on $\partial \Omega$. The singularity remains exactly that of Φ . The two requirements

fix the form of the correction. Adding a harmonic function is the one modification that leaves the equation and the singularity at x untouched, and the boundary values are the only thing it can change. The corrector exists: for interior x one has $|y - x| \geq \text{dist}(x, \partial\Omega) > 0$ on $\partial\Omega$, so the boundary values $\Phi(\cdot - x)$ are smooth, and the Dirichlet problem (3.64) has a unique solution on a smooth bounded domain. Uniqueness is the maximum principle of Section 3.2.

Three properties of the problem make this definition legitimate, and it breaks down when one of them fails. The first is linearity. Since $-\Delta$ is linear,

$$-\Delta_y G = -\Delta_y \Phi + \Delta_y \phi^x = \delta_x + 0, \quad (3.65)$$

so a solution of the homogeneous equation can be added or subtracted without changing the source. For a nonlinear equation this superposition is not available. The second is the existence of a whole-space fundamental solution. An operator with constant coefficients is invariant under translation, so $\Phi(\cdot - x)$ exists for every x and carries the point source. What remains is a homogeneous problem, whose boundary values can be adjusted. The third is that the Dirichlet problem for Laplace's equation is well-posed in the sense of Section 2.4: every continuous boundary datum has exactly one harmonic extension, by the Poisson kernel on a ball and by the maximum principle for uniqueness. This guarantees that the requirement "cancel Φ on the boundary" can always be met. For an ill-posed boundary problem, such as Cauchy data imposed on an elliptic equation, no such corrector exists. The same three properties give the corrector ψ^y of (3.71) for the heat equation, because the initial-boundary-value problem of Section 3.1.3 is well-posed.

A fourth ingredient decides what the corrector has to do. Green's second formula (3.60) is available because $-\Delta$ is formally self-adjoint. It moves the derivatives from u to the kernel and leaves the boundary term $u \partial_\nu v - v \partial_\nu u$, in which one factor is known and the other is not. The corrector must remove the unknown factor, so it depends on the type of boundary condition. For Dirichlet data $\partial_\nu u$ is unknown, and G must vanish on $\partial\Omega$. For Neumann data u is unknown, and one would need $\partial_\nu G = 0$. That cannot be reached by adding a harmonic function, because integrating $-\Delta G = \delta_x$ over Ω gives $-\int_{\partial\Omega} \partial_\nu G = 1$. The source itself has to be corrected by its mean, which leads to the modified Green's function (3.93) of Section 3.3.6.

The unknown term of (3.62) is removed in two steps. First apply (3.60) to u and the corrector ϕ^x , which is harmonic and smooth on all of $\bar{\Omega}$, so that no ball has to be removed:

$$\int_{\Omega} \phi^x f \, dy = \int_{\partial\Omega} \left(u \frac{\partial \phi^x}{\partial \nu} - \phi^x \frac{\partial u}{\partial \nu} \right) dS. \quad (3.66)$$

Since $\phi^x = \Phi(\cdot - x)$ on $\partial\Omega$, the unknown integral in (3.62) equals

$$\int_{\partial\Omega} \Phi \frac{\partial u}{\partial \nu} \, dS = \int_{\partial\Omega} u \frac{\partial \phi^x}{\partial \nu} \, dS - \int_{\Omega} \phi^x f \, dy, \quad (3.67)$$

which involves only the boundary values $u = g$ and the known function ϕ^x . Second, substitute this into (3.62) and collect the terms with $G = \Phi - \phi^x$. For the solution of (3.61) the result is the representation

$$u(x) = \int_{\Omega} G(x, y) f(y) \, dy - \int_{\partial\Omega} g(y) \frac{\partial G}{\partial \nu_y}(x, y) \, dS(y). \quad (3.68)$$

Homogeneous Dirichlet data $g = 0$ leave only the volume integral: applying A^{-1} is integration against G . Nonzero boundary data are read from the conormal derivative of the same kernel. Finding G explicitly means finding the corrector ϕ^x . On an interval, a half-space, and a ball the corrector is elementary. On a general domain it is itself a Dirichlet problem for Laplace's equation,

which is the problem one was solving; the formula remains the structural description of the inverse, and the matrix of Chapter 5 is what one computes instead.

The maximum principle gives two properties of G . The function $G(x, \cdot)$ is harmonic in $\Omega \setminus \{x\}$, vanishes on $\partial\Omega$, and tends to $+\infty$ at x , so it is positive. For $n \geq 3$ the boundary values of ϕ^x are positive, hence $\phi^x > 0$ and $0 < G(x, y) < \Phi(y - x)$: the boundary lowers the potential of a point source. A positive function that vanishes on $\partial\Omega$ decreases in the outward direction, $\partial_\nu G \leq 0$, so the weight $-\partial G/\partial \nu_y$ that multiplies g in (3.68) is nonnegative. The harmonic extension of g is therefore a weighted average of the boundary values. The weights sum to one, because $u \equiv 1$ solves (3.61) with $f = 0$ and $g = 1$. The Poisson kernels of the half-space and the ball below are these weights.

The construction has a physical reading. The function $\Phi(y - x)$ is the potential of a unit point charge at x in free space. If the boundary is a grounded conductor, the potential must vanish there, and the surface acquires an induced charge. The field of that induced charge is $-\phi^x$. For a half-space and for a ball the induced charge can be replaced by one image charge at the mirror point $\tilde{x}$, which is the corrector found below.

A one-dimensional case shows all the steps. On $\Omega = (0, 1)$ the fundamental solution of $-d^2/dy^2$ is $\Phi(y) = -|y|/2$, because $(|y|)'' = 2\delta$. The corrector is the linear function

$$\phi^x(y) = xy - \frac{x+y}{2}. \quad (3.69)$$

It is harmonic, since $\phi^{x''} = 0$. At the endpoints $\phi^x(0) = -x/2 = \Phi(0 - x)$ and $\phi^x(1) = (x - 1)/2 = \Phi(1 - x)$. Therefore

$$G(x, y) = -\frac{|y - x|}{2} - xy + \frac{x + y}{2} = \min\{x, y\}(1 - \max\{x, y\}), \quad (3.70)$$

which is (3.75) with $a = 0$ and $b = 1$. The corner of $-|y - x|/2$ carries the source δ_x , and the linear corrector tilts the graph until it vanishes at both ends.

The same correction applies to the heat equation. On a bounded spatial domain the kernel in (3.28) is replaced by

$$G(x, y, t) = \Gamma(x - y, t) - \psi^y(x, t), \quad (3.71)$$

where, for each frozen source point y and for $t > 0$, the corrector ψ^y solves the heat equation and equals $\Gamma(x - y, t)$ for x on the lateral boundary. Then G vanishes on that boundary, and the representation of an initial-boundary-value problem is (3.28) with Γ replaced by G . The construction is the parabolic form of (3.63).

3.3.2 The interval

On a bounded interval the corrector is a linear function, and G can be written in closed form. Consider

$$-u''(x) = f(x), \quad a < x < b, \quad u(a) = u(b) = 0. \quad (3.72)$$

The Green's function $G(x, \xi)$ satisfies $-\partial_{xx}G = \delta(x - \xi)$ together with $G(a, \xi) = G(b, \xi) = 0$. Away from $x = \xi$ the general solution is linear. Vanishing at the endpoints forces

$$G(x, \xi) = \begin{cases} A(x - a), & a \leq x \leq \xi, \\ B(b - x), & \xi \leq x \leq b. \end{cases} \quad (3.73)$$

Continuity at $x = \xi$ gives $A(\xi - a) = B(b - \xi)$. The derivative jumps: integrating $-\partial_{xx}G = \delta(x - \xi)$ across ξ yields $\partial_x G(\xi-) - \partial_x G(\xi+) = 1$, so $A - (-B) = 1$. Solving the two algebraic equations,

$$A = \frac{b - \xi}{b - a}, \quad B = \frac{\xi - a}{b - a}, \quad (3.74)$$

and therefore

$$G(x, \xi) = \frac{(\min\{x, \xi\} - a)(b - \max\{x, \xi\})}{b - a}. \quad (3.75)$$

The solution of (3.72) is $u(x) = \int_a^b G(x, \xi) f(\xi) d\xi$. Figure 3.3 shows the graph for one source point: two straight pieces meeting at a corner. The corner is the jump of the derivative. The function is symmetric in (x, ξ) , which is the one-dimensional case of Section 3.3.4.

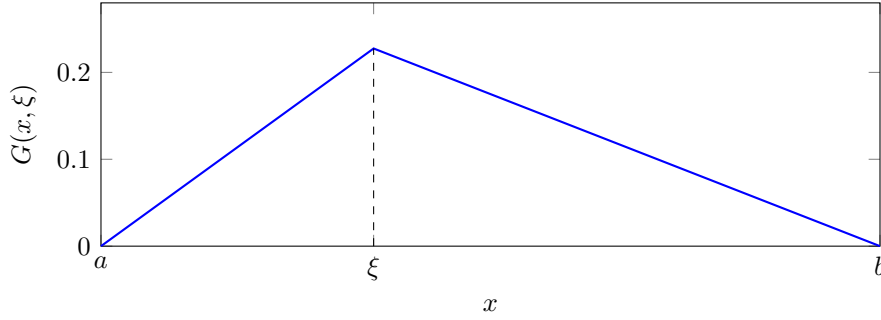


Figure 3.3: The Green's function (3.75) on $(0, 1)$ for $\xi = 0.35$. The height of the corner is $\xi(1 - \xi) = 0.2275$. The slope jumps by -1 across ξ , which is $-\partial_{xx}G = \delta(x - \xi)$.

Section 5.7 replaces (3.72) on $(0, 1)$ by the linear system $A\mathbf{u} = \mathbf{f}$ with the tridiagonal matrix (5.76). That matrix is a consistent discretization of $-d^2/dx^2$, and its inverse samples the kernel: if $h = 1/(n + 1)$ and $x_i = ih$, then

$$(A^{-1})_{ij} = h G(x_i, x_j). \quad (3.76)$$

The factor h is the quadrature weight in $u(x_i) = \int_0^1 G(x_i, \xi) f(\xi) d\xi$. Thus the quantum linear-system problem for this Poisson equation is a discrete reading of integration against (3.75).

3.3.3 The half-space and the ball

On the half-space $\mathbb{R}_+^n = \{x : x_n > 0\}$ the corrector is a reflection. Write $\tilde{x} = (x', -x_n)$ for the mirror image of $x = (x', x_n)$ across the plane $x_n = 0$. For y in the closed half-space and x in the open half-space the point $\tilde{x}$ lies outside $\mathbb{R}_+^n$, so $y \mapsto \Phi(y - \tilde{x})$ is harmonic in y throughout the half-space. On the boundary plane $y_n = 0$ one has $|y - x| = |y - \tilde{x}|$, and since Φ depends only on distance,

$$G(x, y) = \Phi(y - x) - \Phi(y - \tilde{x}) \quad (3.77)$$

vanishes for $y_n = 0$. This is (3.63) with $\phi^x(y) = \Phi(y - \tilde{x})$.

The Poisson kernel is the boundary integrand in (3.68). The outward unit normal on $\{y_n = 0\}$ points out of the half-space, $\nu = -e_n$, so $\partial/\partial\nu = -\partial/\partial y_n$. From (3.48),

$$\frac{\partial}{\partial y_n} \Phi(y - x) = -\frac{1}{n\alpha_n} \frac{y_n - x_n}{|y - x|^n}. \quad (3.78)$$

At $y_n = 0$ the two distances agree and $\tilde{x}_n = -x_n$, so

$$\frac{\partial}{\partial y_n} \Phi(y - x) = \frac{x_n}{n\alpha_n |x - y|^n}, \quad \frac{\partial}{\partial y_n} \Phi(y - \tilde{x}) = -\frac{x_n}{n\alpha_n |x - y|^n}. \quad (3.79)$$

Subtracting these contributions gives

$$\frac{\partial G}{\partial y_n}(x, y) \Big|_{y_n=0} = \frac{2x_n}{n\alpha_n |x - y|^n}. \quad (3.80)$$

For $\Delta u = 0$ in the half-space with $u(y', 0) = g(y')$, formula (3.68) with $f = 0$ therefore gives

$$u(x) = \int_{\mathbb{R}^{n-1}} \frac{2x_n}{n\alpha_n |x - (y', 0)|^n} g(y') dy'. \quad (3.81)$$

The integrand's coefficient is the Poisson kernel of the half-space. The same expression holds for $n = 2$ and for $n \geq 3$, because both cases share the gradient (3.48).

On the ball the corrector is a reflection in the sphere rather than in a plane. For the unit ball and $x \neq 0$ the inverse point $\tilde{x} = x/|x|^2$ lies outside the ball, and $|\tilde{x}| = 1/|x|$. If $|y| = 1$,

$$|x|^2 |y - \tilde{x}|^2 = |x|^2 \left(1 - 2y \cdot \frac{x}{|x|^2} + \frac{1}{|x|^2} \right) = |x|^2 - 2y \cdot x + 1 = |x - y|^2, \quad (3.82)$$

so $|x| |y - \tilde{x}| = |y - x|$. The function $y \mapsto \Phi(|x|(y - \tilde{x}))$ is harmonic inside the ball, because its only singularity is at $\tilde{x}$. On the sphere the distances agree, so for $n \geq 3$

$$\Phi(|x|(y - \tilde{x})) = \frac{1}{n(n-2)\alpha_n (|x| |y - \tilde{x}|)^{n-2}} = \frac{1}{n(n-2)\alpha_n |y - x|^{n-2}} = \Phi(y - x), \quad (3.83)$$

while for $n = 2$ the same identity gives $\log(|x| |y - \tilde{x}|) = \log |y - x|$. Hence

$$G(x, y) = \Phi(y - x) - \Phi(|x|(y - \tilde{x})) \quad (3.84)$$

is the Green's function of the unit ball. The boundary integral (3.68) reduces to the Poisson kernel of a ball of radius r ,

$$K(x, y) = \frac{r^2 - |x|^2}{n\alpha_n r |x - y|^n}, \quad u(x) = \int_{\partial B_r} K(x, y) g(y) dS(y), \quad (3.85)$$

for $\Delta u = 0$ in B_r and $u = g$ on ∂B_r . The half-space formula (3.81) is the limit of (3.85) under dilation. Both are the same correction (3.63): subtract a harmonic function that matches Φ on the boundary. On a general domain the corrector is the Dirichlet problem (3.64), and the representation (3.68) is realized numerically by the matrix A^{-1} .

3.3.4 Symmetry

The kernel is symmetric,

$$G(x, y) = G(y, x), \quad x, y \in \Omega, \quad x \neq y. \quad (3.86)$$

To see it, fix two distinct points and set $v(z) = G(x, z)$, $w(z) = G(y, z)$. Both vanish on $\partial\Omega$, and both are harmonic away from their own singularities. Apply (3.60) on Ω with small balls about x and about y removed. The integral over $\partial\Omega$ vanishes because $v = w = 0$ there. The integrals over the two small spheres reduce, by the same flux computation as for Φ , to $w(x)$ and $v(y)$. They cancel only if $w(x) = v(y)$, which is (3.86). The one-dimensional formula (3.75) is visibly symmetric. The matrix (5.76) is symmetric for the same reason: it is a discretization of an operator whose inverse kernel satisfies (3.86), and the stencil itself was chosen so that the discrete operator remains self-adjoint.

3.3.5 Linear response

The kernel G is the response of the solution to a source. For $-\Delta u = f$ in Ω with $u = 0$ on $\partial\Omega$, formula (3.68) reduces to

$$u(x) = \int_{\Omega} G(x, y) f(y) dy. \quad (3.87)$$

A change of the source changes the field by the same kernel,

$$\delta u(x) = \int_{\Omega} G(x, y) \delta f(y) dy. \quad (3.88)$$

The value $G(x, y)$ is the solution at x produced by a unit source at y . The heat representation (3.28) has the same form with a causal kernel: a source acting at (y, s) contributes at (x, t) through $\Gamma(x - y, t - s)$ when $t > s$.

On the grid of Section 5.7 one physical reading of the field is a linear functional. With homogeneous Dirichlet data, (3.76) converts that functional into the kernel,

$$\sum_i c_i u_i = \sum_{i,j} c_i (A^{-1})_{ij} f_j = \sum_{i,j} c_i h G(x_i, x_j) f_j. \quad (3.89)$$

Section 1.4 asks a quantum algorithm for this scalar—a voltage, a flux, or an overlap—because recovering every entry of $\mathbf{u}$ repeats the algorithm once per unknown.

The same convolution, when the source is a term in a Hamiltonian, is the Kubo formula [31]. Let $H(t) = H_0 - F(t)B$, where the amplitude F is a classical function of time and B is the operator it couples to, and let A be the observable that is read. In the units $\hbar = 1$ of the factor e^{-itH} in Table 1.1, the Heisenberg operator generated by H_0 is $A_H(t) = e^{itH_0} A e^{-itH_0}$, and likewise for B . An equilibrium state ρ_0 commutes with H_0 . Pass to the interaction picture, in which these are the evolved operators and the state carries the perturbation. The von Neumann equation $i\partial_t \rho_I = [V_I, \rho_I]$ with $V = -FB$, kept to first order in F , integrates to

$$\rho_I(t) = \rho_0 + i \int_{-\infty}^t F(s) [B_H(s), \rho_0] ds. \quad (3.90)$$

The deviation of $\langle A \rangle = \text{Tr}(\rho A)$ is $\text{Tr}((\rho_I(t) - \rho_0) A_H(t))$. Cyclicity of the trace and the commutation of ρ_0 with H_0 move the operators into one expectation that depends on the time difference only:

$$\text{Tr}([B_H(s), \rho_0] A_H(t)) = \langle [A_H(t), B_H(s)] \rangle_0 = \langle [A_H(t - s), B] \rangle_0, \quad (3.91)$$

where $\langle \cdot \rangle_0 = \text{Tr}(\rho_0 \cdot)$. Therefore

$$\delta \langle A \rangle(t) = \int_{-\infty}^{\infty} \chi_{AB}(t - s) F(s) ds, \quad \chi_{AB}(t) = i \theta(t) \langle [A_H(t), B] \rangle_0. \quad (3.92)$$

The step $\theta(t)$ equals 1 for $t > 0$ and equals 0 for $t < 0$, so the response at time t uses only the source at earlier times. That is the causality of the heat kernel and of the domain of dependence $|y - x| \leq ct$. The retarded correlator $G_{AB}^R(t) = -i \theta(t) \langle [A_H(t), B] \rangle_0$ is the operator form of the kernel in (3.88), and (3.92) says $\chi_{AB} = -G_{AB}^R$. Formula (3.90) is the linear term in F . Chapter 9 leaves aside the expansion of G^R in a two-body interaction.

3.3.6 Neumann data and the modified Green's function

Replace the Dirichlet condition by a Neumann condition and the construction fails. On the interval consider $-u'' = f$ on $(0, 1)$ with $u'(0) = u'(1) = 0$. Integrating the equation gives $0 = -[u']_0^1 = \int_0^1 f \, dx$, so a solution exists only if $\int f = 0$, and then u is determined only up to an additive constant. The recipe for G breaks in the same way: a function with $-G'' = \delta(x - \xi)$ and $G'(0) = G'(1) = 0$ would give $0 = -[G']_0^1 = \int \delta = 1$. There is no Green's function. In $\mathbb{R}^n$ the divergence theorem gives $\int_\Omega f = -\int_{\partial\Omega} \partial_\nu u = 0$, the compatibility condition of Section 2.3, and the constants span the kernel of the operator. This is the continuous counterpart of the singular matrix of Section 5.8.2, whose null vector is $\mathbf{1}$.

The remedy is to subtract the mean of the source. The modified Green's function satisfies

$$-\Delta_y G(x, y) = \delta_x(y) - \frac{1}{|\Omega|} \quad \text{in } \Omega, \quad \partial_\nu G = 0 \quad \text{on } \partial\Omega, \quad \int_\Omega G(x, y) \, dy = 0. \quad (3.93)$$

The right side now has zero mean, so the problem is solvable, and the last condition removes the free constant. On $(0, 1)$ the solution of $-G'' = \delta(x - \xi) - 1$ with $G'(0) = 0$ is $G'(x) = x - \mathbf{1}_{x > \xi}$, which also satisfies $G'(1) = 0$. Integrating once more and imposing $\int_0^1 G \, dx = 0$,

$$G(x, \xi) = \frac{x^2 + \xi^2}{2} - \max\{x, \xi\} + \frac{1}{3}. \quad (3.94)$$

It is symmetric in (x, ξ) , and its eigenfunction expansion is $G(x, \xi) = \sum_{k \geq 1} \frac{2 \cos(k\pi x) \cos(k\pi \xi)}{(k\pi)^2}$, the cosine modes with eigenvalues $(k\pi)^2$ and without the mode $k = 0$. For a source with $\int f = 0$ the solution is $u(x) = \int G(x, y) f(y) \, dy$ plus a constant, because the added term $1/|\Omega|$ integrates against f to zero. The matrix analogue is the pseudoinverse: G is the kernel of A^+ on $\mathbf{1}^\perp$, the right-hand side must lie in the range of A , and the condition number is measured by the smallest nonzero eigenvalue, which is π^2 in the continuum and λ_1 of (5.103) on the grid. The quantum linear-system solvers of Chapter 9 invert only the nonzero singular values, so the pure Neumann problem is a legitimate instance after the source has been projected onto $\mathbf{1}^\perp$.

3.3.7 The variational form

The symmetry of G and of the matrix A has a variational counterpart. For $-\Delta u = f$ in Ω with $u = 0$ on $\partial\Omega$ define the energy

$$J(v) = \int_\Omega \left(\frac{1}{2} |\nabla v|^2 - f v \right) \, dx, \quad v = 0 \text{ on } \partial\Omega. \quad (3.95)$$

The solution u minimizes J . For any admissible w with $w = 0$ on $\partial\Omega$,

$$J(u + w) = J(u) + \int_\Omega (\nabla u \cdot \nabla w - f w) \, dx + \frac{1}{2} \int_\Omega |\nabla w|^2 \, dx. \quad (3.96)$$

Green's first formula (3.59) with the boundary term absent gives $\int \nabla u \cdot \nabla w = -\int w \Delta u = \int f w$, so the middle term vanishes and $J(u + w) = J(u) + \frac{1}{2} \int |\nabla w|^2 \geq J(u)$, with equality only for $\nabla w = 0$, hence $w = 0$. The solution is the unique minimizer, which is another proof of uniqueness. Conversely, a minimizer satisfies $-\Delta u = f$, the Euler–Lagrange equation, so Poisson's equation is a minimum problem. Dropping the boundary constraint leads to the Neumann condition as a natural boundary condition: $J(v + \text{const}) = J(v) - \text{const} \int f$ is bounded below only if $\int f = 0$, the compatibility condition above.

On the grid the energy is $J(\mathbf{v}) = \frac{1}{2}\mathbf{v}^\top A\mathbf{v} - \mathbf{b}^\top \mathbf{v}$ for a symmetric positive definite A , whose minimizer solves $A\mathbf{v} = \mathbf{b}$. The finite element method of Section 5.11 minimizes J over a finite-dimensional subspace V_h , and the identity $J(v) - J(u) = \frac{1}{2}\|\nabla(v - u)\|^2$ shows that the Galerkin solution is the best approximation of u in the energy norm. Variational quantum solvers, Chapter 16, likewise replace inversion by the minimization of a cost over parameters, and the energy J is one such cost for a symmetric positive definite system.

3.4 The wave equation

3.4.1 d'Alembert's formula

The one-dimensional wave equation

$$u_{tt} = c^2 u_{xx}, \quad x \in \mathbb{R}, \quad t > 0, \quad (3.97)$$

with $c > 0$, asks for Cauchy data $u(x, 0) = \phi(x)$ and $u_t(x, 0) = \psi(x)$. The characteristic coordinates of Section 2.5.3 are $\xi = x + ct$ and $\eta = x - ct$. In these coordinates

$$\partial_x = \partial_\xi + \partial_\eta, \quad \partial_t = c\partial_\xi - c\partial_\eta. \quad (3.98)$$

A second differentiation gives

$$\begin{aligned} \partial_{xx} &= \partial_{\xi\xi} + 2\partial_{\xi\eta} + \partial_{\eta\eta}, \\ \partial_{tt} &= c^2\partial_{\xi\xi} - 2c^2\partial_{\xi\eta} + c^2\partial_{\eta\eta}, \end{aligned} \quad (3.99)$$

and therefore

$$u_{tt} - c^2 u_{xx} = -4c^2 u_{\xi\eta}. \quad (3.100)$$

Equation (3.97) is $u_{\xi\eta} = 0$. Integrating in ξ shows that u_η is independent of ξ , and integrating again shows that

$$u(x, t) = F(x + ct) + G(x - ct) \quad (3.101)$$

for two profiles F and G of a single variable. The initial conditions become $F(x) + G(x) = \phi(x)$ and $c(F'(x) - G'(x)) = \psi(x)$. Integrate the second identity from 0 to s and solve the linear system:

$$\begin{aligned} F(s) &= \frac{\phi(s)}{2} + \frac{1}{2c} \int_0^s \psi(\sigma) d\sigma + A, \\ G(s) &= \frac{\phi(s)}{2} - \frac{1}{2c} \int_0^s \psi(\sigma) d\sigma + B, \end{aligned} \quad (3.102)$$

with $A + B = 0$ because $F(0) + G(0) = \phi(0)$. Substituting $s = x + ct$ and $s = x - ct$, the constants cancel and the two integrals of ψ combine into one interval. The result is d'Alembert's formula

$$u(x, t) = \frac{\phi(x + ct) + \phi(x - ct)}{2} + \frac{1}{2c} \int_{x-ct}^{x+ct} \psi(s) ds. \quad (3.103)$$

The value at (x, t) depends only on the data in the interval $[x - ct, x + ct]$. That interval is the domain of dependence. Its endpoints travel along the characteristics $x \pm ct = \text{const}$ at speed c , which is the finite propagation speed already read off from the type in Section 2.5.3. The first term propagates the displacement, half to the right and half to the left. The integral propagates the initial velocity, uniformly across the interval that the two waves have swept. A corner in ϕ remains a corner, transported along the two families of lines.

3.4.2 The half-line

On the half-line $x > 0$ with a homogeneous Dirichlet condition $u(0, t) = 0$, the same formula applies after the data have been extended oddly. Set $\tilde{\phi}(-x) = -\phi(x)$ and $\tilde{\psi}(-x) = -\psi(x)$ for $x > 0$, and $\tilde{\phi}(0) = \tilde{\psi}(0) = 0$. The odd extension of the solution of the whole-line problem stays odd for all later times, because (3.103) preserves oddness, and an odd continuous function vanishes at $x = 0$. Restricting to $x > 0$ gives the solution of the half-line problem. Explicitly, the backward characteristic from (x, t) meets the initial line before it meets the boundary when $x \geq ct$, and the formula is (3.103) with the original data. When $0 \leq x \leq ct$ the left-going ray has reflected, the odd extension supplies a minus sign, and

$$u(x, t) = \frac{\phi(ct + x) - \phi(ct - x)}{2} + \frac{1}{2c} \int_{ct-x}^{ct+x} \psi(s) ds. \quad (3.104)$$

Reflection at a Dirichlet end reverses the sign of the incoming wave. A Neumann condition $u_x(0, t) = 0$ is the same construction with even extensions, and the reflected wave keeps its sign.

3.4.3 Energy and finite propagation speed

The wave equation conserves an energy, and this is the structural difference from the heat equation. For $u_{tt} = c^2 \Delta u$ on $\mathbb{R}^n$ with data of compact support, put

$$E(t) = \frac{1}{2} \int_{\mathbb{R}^n} (u_t^2 + c^2 |\nabla u|^2) dx. \quad (3.105)$$

The first term is the kinetic energy of the motion and the second is the elastic energy stored in the stretching of the string or membrane. Differentiate under the integral, use $\nabla \cdot (u_t \nabla u) = \nabla u_t \cdot \nabla u + u_t \Delta u$, and apply the divergence theorem (3.56):

$$\begin{aligned} E'(t) &= \int_{\mathbb{R}^n} (u_t u_{tt} + c^2 \nabla u \cdot \nabla u_t) dx \\ &= \int_{\mathbb{R}^n} u_t (u_{tt} - c^2 \Delta u) dx + c^2 \int_{\mathbb{R}^n} \nabla \cdot (u_t \nabla u) dx = 0. \end{aligned} \quad (3.106)$$

The first integral vanishes by the equation and the second by decay at infinity (or by $u = 0$, hence $u_t = 0$, or $\partial_\nu u = 0$ on the boundary of a bounded domain). Compare (3.34): for the heat equation the same multiplication gives $\frac{1}{2} \frac{d}{dt} \int w^2 = - \int |\nabla w|^2 \leq 0$, an inequality, so the heat flow contracts and forgets its data. The wave flow conserves E and is reversible, $t \mapsto -t$ being a symmetry of the equation. In the language of Table 1.1 the heat equation is the contraction e^{tA} and the wave equation, written as a first-order system for the pair (u, u_t) , is the norm-preserving group e^{-itH} with $H = c\sqrt{-\Delta}$, and E is the squared norm. In the sine basis of (3.132) every mode keeps its own energy.

Two consequences follow without any formula. If u_1 and u_2 solve the same Cauchy problem, then $u = u_1 - u_2$ has $E(0) = 0$, hence $E(t) = 0$ for all t , hence $u_t = 0$ and $\nabla u = 0$; since $u(\cdot, 0) = 0$, one gets $u \equiv 0$. The Cauchy problem has at most one solution in every dimension. The second consequence is finite propagation speed, which can be proved in the same way on a shrinking ball. For $x_0 \in \mathbb{R}^n$, $R > 0$, and $0 \leq t \leq R/c$ let

$$e(t) = \frac{1}{2} \int_{B_{R-ct}(x_0)} (u_t^2 + c^2 |\nabla u|^2) dx. \quad (3.107)$$

The domain shrinks at speed c , which contributes a boundary term from the moving sphere, and the integrand is handled as in (3.106) with the divergence theorem on the ball:

$$e'(t) = \int_{\partial B_{R-ct}(x_0)} \left(c^2 u_t \partial_\nu u - \frac{c}{2} (u_t^2 + c^2 |\nabla u|^2) \right) dS \leq 0. \quad (3.108)$$

The integrand is nonpositive because $c^2 |u_t \partial_\nu u| \leq c^2 |u_t| |\nabla u| \leq \frac{c}{2} (u_t^2 + c^2 |\nabla u|^2)$, which is the inequality $ab \leq (a^2 + b^2)/2$ with $a = u_t$ and $b = c |\nabla u|$. If u and u_t vanish on $B_R(x_0)$ at $t = 0$, then $e(0) = 0$, so $e \equiv 0$, and u vanishes on the whole cone $\{|x - x_0| < R - ct\}$. Taking $R = ct_0$ shows that $u(x_0, t_0)$ is determined by the data on $\overline{B_{ct_0}(x_0)}$. This is the domain of dependence of Section 3.4.1 in any dimension.

3.4.4 Higher dimensions: spherical means

In one dimension the factorization $\partial_t^2 - c^2 \partial_x^2 = (\partial_t - c \partial_x)(\partial_t + c \partial_x)$ gives the general solution $F(x + ct) + G(x - ct)$. In $\mathbb{R}^n$ with $n \geq 2$ no such splitting into two transport equations exists, and the plan is to reduce to one space variable by averaging over spheres. Fix $x \in \mathbb{R}^n$ and write $\langle f \rangle_{x,r} = \frac{1}{|\partial B_r(x)|} \int_{\partial B_r(x)} f dS$ for the mean of f over the sphere of radius r about x , where $|\partial B_r| = n \alpha_n r^{n-1}$ as in Section 3.2. Define

$$U(r, t) = \langle u(\cdot, t) \rangle_{x,r}, \quad G(r) = \langle \phi \rangle_{x,r}, \quad H(r) = \langle \psi \rangle_{x,r}. \quad (3.109)$$

The center is suppressed in the notation. Since u is continuous, $u(x, t) = \lim_{r \rightarrow 0} U(r, t)$, so the whole problem is to find U . The computation of (3.53) did not use the equation, only the divergence theorem, so it also holds here:

$$r^{n-1} U_r(r, t) = \frac{1}{n \alpha_n} \int_{B_r(x)} \Delta u(y, t) dy. \quad (3.110)$$

Differentiate in r . The derivative of an integral over B_r is the integral over ∂B_r of the integrand, so

$$\partial_r (r^{n-1} U_r) = \frac{1}{n \alpha_n} \int_{\partial B_r(x)} \Delta u dS = r^{n-1} \langle \Delta u \rangle_{x,r} = \frac{r^{n-1}}{c^2} U_{tt}. \quad (3.111)$$

The last equality uses $\Delta u = c^{-2} u_{tt}$ and the fact that the time derivative commutes with the spatial mean. Expanding the left side and dividing by r^{n-1} gives the Euler–Poisson–Darboux equation

$$U_{tt} = c^2 \left(U_{rr} + \frac{n-1}{r} U_r \right), \quad U(r, 0) = G(r), \quad U_t(r, 0) = H(r). \quad (3.112)$$

This is the radial wave equation, and it is an equation in the two variables (r, t) whatever the dimension. The dimension enters only through the coefficient $(n-1)/r$ of the first-order term. For $n = 1$ the “sphere” $\partial B_r(x)$ is the pair of points $x \pm r$, $U = \frac{1}{2}(u(x+r, t) + u(x-r, t))$, the lower-order term is absent, and (3.112) is the one-dimensional wave equation, solved by d’Alembert’s formula. The remaining work is to remove the term $(n-1)U_r/r$ by a change of unknown.

3.4.5 Three dimensions: Kirchhoff’s formula

For $n = 3$ put $\tilde{U} = rU$. Then $\tilde{U}_{rr} = rU_{rr} + 2U_r = r(U_{rr} + \frac{2}{r}U_r)$, and by (3.112) this equals rU_{tt}/c^2 . Hence

$$\tilde{U}_{tt} = c^2 \tilde{U}_{rr}, \quad r > 0, \quad \tilde{U}(r, 0) = \tilde{G}(r) = rG(r), \quad \tilde{U}_t(r, 0) = \tilde{H}(r) = rH(r). \quad (3.113)$$

Because U is bounded as $r \rightarrow 0$, also $\tilde{U}(0, t) = 0$. This is the wave equation on the half-line with a Dirichlet condition, solved in (3.104). For $0 < r < ct$,

$$\tilde{U}(r, t) = \frac{\tilde{G}(ct + r) - \tilde{G}(ct - r)}{2} + \frac{1}{2c} \int_{ct-r}^{ct+r} \tilde{H}(s) ds. \quad (3.114)$$

Divide by r and let $r \rightarrow 0$. The first quotient tends to $\tilde{G}'(ct)$ and the average of $\tilde{H}$ over the shrinking interval tends to $\tilde{H}(ct)/c$, so

$$u(x, t) = \lim_{r \rightarrow 0} \frac{\tilde{U}(r, t)}{r} = \tilde{G}'(ct) + \frac{1}{c} \tilde{H}(ct). \quad (3.115)$$

Now $\tilde{G}(s) = s\langle\phi\rangle_{x,s}$, so $\tilde{G}'(ct) = \frac{1}{c}\partial_t[ct\langle\phi\rangle_{x,ct}] = \partial_t(t\langle\phi\rangle_{x,ct})$, and $\tilde{H}(ct)/c = t\langle\psi\rangle_{x,ct}$. The sphere of radius ct has area $4\pi c^2 t^2$. Therefore

$$\begin{aligned} u(x, t) &= \partial_t(t\langle\phi\rangle_{x,ct}) + t\langle\psi\rangle_{x,ct} \\ &= \frac{\partial}{\partial t} \left(\frac{1}{4\pi c^2 t} \int_{|y-x|=ct} \phi(y) dS(y) \right) + \frac{1}{4\pi c^2 t} \int_{|y-x|=ct} \psi(y) dS(y), \end{aligned} \quad (3.116)$$

which is Kirchhoff's formula. The derivation shows that a C^2 solution must be given by this expression. That the expression is in turn a solution is checked by differentiating it, which is routine once the formula is written in the following form. Substitute $y = x + ctz$ with $|z| = 1$ in the ϕ term and differentiate in t under the integral:

$$u(x, t) = \langle \phi(y) + \nabla \phi(y) \cdot (y - x) + t\psi(y) \rangle_{x,ct}. \quad (3.117)$$

The value at (x, t) is an average over the sphere of radius ct , of the data and of the first derivative of ϕ along the radius. Unlike d'Alembert's formula this requires derivatives of the data: for $u \in C^2$ one needs $\phi \in C^3$ and $\psi \in C^2$. The loss of a derivative is the focusing of spherical waves, since the average over a sphere that shrinks to a point near the origin is sensitive to the data's gradient.

3.4.6 Odd dimensions

The substitution $\tilde{U} = rU$ worked because $(rU)_{rr} = r(U_{rr} + \frac{2}{r}U_r)$. For $n = 5$ the operator that does the same is $\tilde{U} = \frac{1}{r}\partial_r(r^3U) = 3rU + r^2U_r$. Its two second derivatives are computed from $U_{tt} = c^2(U_{rr} + \frac{4}{r}U_r)$. In space,

$$\tilde{U}_{rr} = r^2U_{rrr} + 7rU_{rr} + 8U_r. \quad (3.118)$$

In time, $\tilde{U}_{tt} = 3rU_{tt} + r^2\partial_r U_{tt}$, and inserting (3.112) with its r -derivative $\partial_r U_{tt} = c^2(U_{rrr} + \frac{4}{r}U_{rr} - \frac{4}{r^2}U_r)$ gives

$$\tilde{U}_{tt} = c^2(3rU_{rr} + 12U_r + r^2U_{rrr} + 4rU_{rr} - 4U_r) = c^2(r^2U_{rrr} + 7rU_{rr} + 8U_r). \quad (3.119)$$

The two expressions agree, so $\tilde{U}_{tt} = c^2\tilde{U}_{rr}$, and the half-line argument applies again with $\tilde{G} = \frac{1}{r}\partial_r(r^3G)$ and $\tilde{H} = \frac{1}{r}\partial_r(r^3H)$. The only change in the limit is $\tilde{U}/r \rightarrow 3u(x, t)$, because $\tilde{U} = 3rU + O(r^2)$.

The pattern holds for every odd dimension $n = 2k + 1$. By induction on k , with the same kind of computation as above, the function

$$\tilde{U}(r, t) = \left(\frac{1}{r}\partial_r\right)^{k-1}(r^{2k-1}U(r, t)) \quad (3.120)$$

satisfies $\tilde{U}_{tt} = c^2 \tilde{U}_{rr}$, vanishes at $r = 0$, and obeys $\tilde{U}/r \rightarrow \gamma_n u(x, t)$ with $\gamma_n = 1 \cdot 3 \cdot 5 \cdots (n-2)$. Solve the half-line problem, divide by r , and convert $\frac{1}{s} \partial_s$ with $s = ct$ into $\frac{1}{c^2 t} \partial_t$. The speed c then enters only through the radius of the sphere, and

$$u(x, t) = \frac{1}{\gamma_n} \left[\partial_t \left(\frac{1}{t} \partial_t \right)^{\frac{n-3}{2}} \left(t^{n-2} \langle \phi \rangle_{x, ct} \right) + \left(\frac{1}{t} \partial_t \right)^{\frac{n-3}{2}} \left(t^{n-2} \langle \psi \rangle_{x, ct} \right) \right], \quad n \geq 3 \text{ odd.} \quad (3.121)$$

For $n = 3$ the powers of $\frac{1}{t} \partial_t$ are absent, $\gamma_3 = 1$, and (3.121) is (3.116).

3.4.7 Even dimensions: descent

For even n no substitution removes the first-order term in (3.112). The method of descent avoids it: a solution in $\mathbb{R}^n$ is regarded as a solution in $\mathbb{R}^{n+1}$ that does not depend on the last variable. For $n = 2$ set $\bar{u}(x_1, x_2, x_3, t) = u(x_1, x_2, t)$ with data $\bar{\phi}, \bar{\psi}$ defined in the same way. Then $\bar{u}$ solves the three-dimensional wave equation, and (3.116) at $\bar{x} = (x_1, x_2, 0)$ gives $u(x, t)$. The sphere $\{|y - \bar{x}| = ct\}$ consists of two hemispheres, the graphs $y_3 = \pm \gamma(y)$ over the disk $B_{ct}(x) \subset \mathbb{R}^2$, with $\gamma(y) = \sqrt{c^2 t^2 - |y - x|^2}$. The area element of a graph is $\sqrt{1 + |\nabla \gamma|^2} dy$, and

$$1 + |\nabla \gamma|^2 = 1 + \frac{|y - x|^2}{c^2 t^2 - |y - x|^2} = \frac{c^2 t^2}{c^2 t^2 - |y - x|^2}, \quad (3.122)$$

so that $dS = \frac{ct}{\sqrt{c^2 t^2 - |y - x|^2}} dy$ on each hemisphere. Both hemispheres carry the same values of $\bar{\phi}$ and $\bar{\psi}$, and

$$t \langle \bar{\psi} \rangle_{\bar{x}, ct} = \frac{t}{4\pi c^2 t^2} \cdot 2 \int_{B_{ct}(x)} \frac{ct \psi(y)}{\sqrt{c^2 t^2 - |y - x|^2}} dy = \frac{1}{2\pi c} \int_{B_{ct}(x)} \frac{\psi(y)}{\sqrt{c^2 t^2 - |y - x|^2}} dy. \quad (3.123)$$

The ϕ term is ∂_t of the same expression with ψ replaced by ϕ . This is the two-dimensional formula

$$u(x, t) = \frac{1}{2\pi c} \frac{\partial}{\partial t} \int_{|y-x| \leq ct} \frac{\phi(y)}{\sqrt{c^2 t^2 - |y - x|^2}} dy + \frac{1}{2\pi c} \int_{|y-x| \leq ct} \frac{\psi(y)}{\sqrt{c^2 t^2 - |y - x|^2}} dy. \quad (3.124)$$

The same descent from $n + 1$ to n gives every even dimension from the odd formula (3.121) in dimension $n + 1$. The integral is over the solid ball, not over a sphere, and the kernel $(c^2 t^2 - |y - x|^2)^{-1/2}$ is positive in the interior.

3.4.8 Huygens' principle and the wake

The formulas differ in what they say about the support of a pulse. Let ϕ and ψ be supported in a compact set K . In odd dimension $n \geq 3$, (3.121) involves ϕ and ψ only through their values (and finitely many derivatives) on the sphere $|y - x| = ct$. Hence $u(x, t) \neq 0$ is possible only if the sphere meets K , that is, for $\text{dist}(x, K) \leq ct \leq \max_{y \in K} |y - x|$. A disturbance is felt when the front arrives and is gone when the back of the pulse has passed. This is Huygens' principle. A sharp sound signal in air is the physical case, $n = 3$.

In even dimension the integral runs over the ball $|y - x| \leq ct$ with a positive weight, so once the front has reached x the ball keeps containing part of K for all later times and the kernel is not zero there. The disturbance has a tail, a wake that decays like $1/t$ but never vanishes: ripples on

a pond keep ringing after the stone has gone. The same happens for the velocity data ψ in $n = 1$. In d'Alembert's formula (3.103) the term $\frac{1}{2c} \int_{x-ct}^{x+ct} \psi$ is constant once the interval contains $\text{supp } \psi$, but not zero. The displacement data ϕ in one dimension are transported without a tail, which is why Huygens' principle is stated for odd $n \geq 3$ and not for $n = 1$. The principle is a property of the continuous equation in odd dimension $n \geq 3$; the sparse matrix of Chapter 5 that replaces $-\Delta$ does not reproduce it exactly. Numerical dispersion leaves a tail behind the front. The cone itself survives in an effective form: a stencil of fixed width moves information over a bounded number of grid points per time step, and the exact semidiscrete evolution e^{-itH} with a sparse H has only an exponentially small tail outside the cone.

3.4.9 A source term

For $u_{tt} - c^2 \Delta u = f(x, t)$ with $u(x, 0) = u_t(x, 0) = 0$ the solution is again a superposition of homogeneous solutions, as in (3.28). For each $s \geq 0$ let $v(\cdot, \cdot; s)$ solve the homogeneous equation for $t > s$ with $v(x, s; s) = 0$ and $v_t(x, s; s) = f(x, s)$. The forcing at time s is thus a velocity impulse. Set

$$u(x, t) = \int_0^t v(x, t; s) ds. \quad (3.125)$$

Differentiating, $u_t = v(x, t; t) + \int_0^t v_t ds = \int_0^t v_t ds$ because $v(x, t; t) = 0$, and $u_{tt} = v_t(x, t; t) + \int_0^t v_{tt} ds = f(x, t) + c^2 \Delta u$. The equation is satisfied, and the data vanish at $t = 0$. By time translation $v(x, t; s) = w(x, t - s; s)$ where $w(\cdot, \cdot; s)$ solves the homogeneous problem with $\phi = 0$ and $\psi = f(\cdot, s)$. In three dimensions (3.116) gives $w(x, \tau; s) = \tau \langle f(\cdot, s) \rangle_{x, c\tau}$, so with $\tau = t - s$ and the radius $\rho = c\tau$,

$$u(x, t) = \int_0^t (t - s) \langle f(\cdot, s) \rangle_{x, c(t-s)} ds = \frac{1}{4\pi c^2} \int_{|y-x| \leq ct} \frac{f(y, t - |y-x|/c)}{|y-x|} dy. \quad (3.126)$$

This is the retarded potential. The source at y acts on x with the weight $1/(4\pi c^2 |y-x|)$, the Newtonian kernel Φ of (3.47) divided by c^2 , with f evaluated at the earlier time $t - |y-x|/c$ at which a signal travelling at speed c would have left y . For $c \rightarrow \infty$ the delay disappears and (3.126) becomes $c^{-2}(\Phi * f)$ of (3.51), the solution of the stationary equation $-c^2 \Delta u = f$. The causality is the same as in the Kubo response (3.92): only the past of (x, t) contributes.

A bounded domain adds reflections, by the same odd or even extension when the boundary is flat, and by a boundary corrector of the kind in (3.63) when it is not. The formulas of this section are the continuous form of the unitary group e^{-itH} applied to Cauchy data (ϕ, ψ) .

3.5 Series solutions and special functions

On a bounded interval the kernels of the previous sections become series. Separate variables in $u_t = u_{xx}$ on $(0, L)$ with $u(0, t) = u(L, t) = 0$: a product $X(x)T(t)$ satisfies $T'/T = X''/X = -\lambda$. The spatial factor with the Dirichlet conditions is the Sturm–Liouville problem $-X'' = \lambda X$, whose solutions are

$$X_n(x) = \sin\left(\frac{n\pi x}{L}\right), \quad \lambda_n = \left(\frac{n\pi}{L}\right)^2, \quad n = 1, 2, \dots \quad (3.127)$$

The family is orthogonal on $(0, L)$ with $\int_0^L X_n X_m dx = (L/2) \delta_{nm}$. Expanding the initial datum $\phi = \sum a_n X_n$ with

$$a_n = \frac{2}{L} \int_0^L \phi(x) \sin\left(\frac{n\pi x}{L}\right) dx \quad (3.128)$$

and attaching the time factor $e^{-\lambda_n t}$ of the separated equation gives

$$u(x, t) = \sum_{n=1}^{\infty} a_n e^{-\lambda_n t} \sin\left(\frac{n\pi x}{L}\right). \quad (3.129)$$

The series is (3.20) after the heat kernel on the interval has been expanded in the same basis. High modes decay like $e^{-(n\pi/L)^2 t}$, which is the analytic form of the smoothing visible in Figure 3.1.

The orthogonality of the sine modes is a consequence of symmetry, not of the particular functions. Take the inner product $\langle f, g \rangle = \int_0^L f g \, dx$ and $\mathcal{L} = -d^2/dx^2$ on functions with $f(0) = f(L) = 0$. Integrating by parts twice, the boundary terms $f'g$ and $f g'$ vanish, and

$$\langle \mathcal{L}f, g \rangle = \int_0^L f' g' \, dx = \langle f, \mathcal{L}g \rangle, \quad \langle \mathcal{L}f, f \rangle = \int_0^L (f')^2 \, dx > 0 \quad (f \neq 0). \quad (3.130)$$

This is Green's second identity (3.60) in one dimension, and it is the infinite-dimensional version of $A = A^\top$. The finite-dimensional proof carries over word for word. If $\mathcal{L}X_i = \lambda_i X_i$ and $\mathcal{L}X_j = \lambda_j X_j$, then $\lambda_i \langle X_i, X_j \rangle = \langle \mathcal{L}X_i, X_j \rangle = \langle X_i, \mathcal{L}X_j \rangle = \lambda_j \langle X_i, X_j \rangle$, so $\langle X_i, X_j \rangle = 0$ whenever $\lambda_i \neq \lambda_j$, and the eigenvalues are real and positive by the second relation in (3.130). Taking the inner product of $\phi = \sum a_n X_n$ with X_m then isolates $a_m \langle X_m, X_m \rangle$, which is (3.128): the coefficients are orthogonal projections.

What is new in infinite dimension is completeness. A real symmetric matrix has an orthonormal eigenbasis by the spectral theorem. For functions one must also know that no nonzero function is orthogonal to all X_n , so that the series reproduces ϕ . The right setting is the space $L^2(0, L)$ of square-integrable functions, a Hilbert space: a vector space with an inner product, whose norm $\|f\| = \sqrt{\langle f, f \rangle}$ defines a distance, and which is complete, so that every Cauchy sequence converges to an element of the space. Completeness is the property that $\mathbb{R}$ has and $\mathbb{Q}$ lacks, and the one that continuous functions with the L^2 norm lack. In $L^2(0, L)$ the partial sums of the sine series converge to ϕ in norm, and Parseval's identity $\int_0^L \phi^2 \, dx = \frac{L}{2} \sum a_n^2$ holds. The spectral theorem applies to the inverse of $\mathcal{L}$: it is the integral operator with the symmetric kernel (3.75), which is compact, and its eigenfunctions are those of $\mathcal{L}$ with eigenvalues $1/\lambda_n$. The symmetry (3.86) of the Green's function is therefore the reason that an eigenbasis exists. The same argument for $-(pu')' + qu = \lambda wu$ with a weight $w > 0$ gives orthogonality with respect to $\int f g w \, dx$ and produces the Legendre, Bessel, and Hermite families below.

For the quantum algorithms the point is the matrix version. The discretization A of Section 5.7 is real symmetric, so $A = V \Lambda V^\top$ with orthogonal V , and every function of A is $f(A) = V f(\Lambda) V^\top$. For $f(\lambda) = e^{-it\lambda}$ the matrix $f(A)$ is unitary, which is why A can be treated as a Hamiltonian, and for $f(\lambda) = 1/\lambda$ it is the inverse whose condition number $\lambda_{\max}/\lambda_{\min}$ is read from the same spectrum.

The same coefficients solve the other two normal forms. For $-u'' = f$ with $u(0) = u(L) = 0$ and $f = \sum f_n X_n$,

$$u(x) = \sum_{n=1}^{\infty} \frac{f_n}{\lambda_n} \sin\left(\frac{n\pi x}{L}\right). \quad (3.131)$$

Division by λ_n is integration against the Green's function (3.75), written in the eigenbasis: the integral operator with kernel G is diagonal, with eigenvalues $1/\lambda_n$. For the wave equation $u_{tt} = c^2 u_{xx}$ with the same boundary conditions, the time factor is an oscillation of frequency $c\sqrt{\lambda_n}$,

$$u(x, t) = \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi c t}{L} + b_n \sin \frac{n\pi c t}{L} \right) \sin\left(\frac{n\pi x}{L}\right), \quad (3.132)$$

where a_n are the sine coefficients of ϕ and $b_n = \frac{L}{n\pi c}$ times the sine coefficients of ψ . The series is d'Alembert's formula summed in the Dirichlet basis. Every term oscillates at frequency $n\pi c/L$.

Section 5.7 discretizes $-u''$ on $(0, 1)$ by the matrix (5.76). Its eigenvectors are the grid samples of (3.127), and its eigenvalues are $(4/h^2) \sin^2(k\pi h/2)$ with $h = 1/(n+1)$. As $h \rightarrow 0$ with k fixed, $\sin(k\pi h/2) \sim k\pi h/2$, so the discrete eigenvalues tend to $(k\pi)^2 = \lambda_k$. The discrete sine transform of Section 5.10 is the finite section of the expansion used in (3.131), (3.129), and (3.132). Applying A^{-1} , e^{-tA} , or $\cos(t\sqrt{A})$ on that basis is the series, truncated to the grid.

The three kernels and the three series are one construction. On the whole space the construction is a convolution. On an interval it is a multiplier $\frac{1}{\lambda_n}$, $e^{-\lambda_n t}$, or $e^{\pm ic\sqrt{\lambda_n} t}$ in the sine basis. On a grid the multiplier is the same function of the eigenvalues of a sparse matrix. The chapters that follow compute that function of the matrix. The sine series is the constant-coefficient case on an interval. The subsections that follow are the separated factors of a point source, the Coulomb potential, a quadratic well, and a circular membrane: a power series $y = \sum_{k \geq 0} a_k x^k$ solves the ordinary equation that remains once the ratios a_{k+2}/a_k or a_{k+1}/a_k are fixed, and an integer index terminates the chain or sums it in closed form [32, 30]. The inner product that makes those solutions orthogonal, and the least-squares use of that product, are in Section 5.4. Section 3.6 records the equations for which superposition stops; Chapter 15 returns to them as a lift back to a linear evolution.

3.5.1 Legendre polynomials

A point mass and a point charge produce the same kernel. In three dimensions $\Phi(x) = 1/(4\pi|x|)$, so the potential (3.51) is a constant multiple of $\int f(y)/|x-y| dy$. The gravitational potential of a mass density and the electrostatic potential of a charge density are that integral, multiplied by $-G$ and by $1/(4\pi\epsilon_0)$ respectively. For an observation point with $r = |x|$ and a source point x' with $r' = |x'| < r$,

$$\frac{1}{|x-x'|} = \frac{1}{r\sqrt{1-2\mu t+t^2}}, \quad t = \frac{r'}{r}, \quad \mu = \frac{x \cdot x'}{rr'}. \quad (3.133)$$

The expansion of the square root in powers of t defines a polynomial $p_n(\mu)$ of degree n as the coefficient of t^n . The kernel is therefore the sum

$$\frac{1}{|x-x'|} = \sum_{n=0}^{\infty} \frac{(r')^n}{r^{n+1}} p_n(\mu). \quad (3.134)$$

A finite collection of point sources is a sum of such kernels, and each term contributes p_n of the angle between the field point and that source. For a density supported inside the sphere of radius r , the same series goes under the integral: the coefficient of r^{-n-1} is the moment $\int (r')^n p_n(\mu) f(x') dx'$. Each term $r^{-n-1} p_n(\cos \theta)$ is harmonic for $r > 0$, so the angular factor satisfies Legendre's equation

$$(1-x^2)y'' - 2xy' + n(n+1)y = 0. \quad (3.135)$$

Substitute $y = \sum_{k \geq 0} a_k x^k$. The coefficient of x^k is

$$(k+2)(k+1)a_{k+2} = (k(k+1) - n(n+1))a_k = (k-n)(k+n+1)a_k. \quad (3.136)$$

If n is a nonnegative integer the factor $k-n$ vanishes at $k=n$, and the even chain or the odd chain stops. The resulting polynomial of degree n is p_n . Rodrigues' formula packages it:

$$p_n(x) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dx^n} [(1-x^2)^n], \quad n \geq 1, \quad (3.137)$$

with $p_0 = 1$, so $p_1 = x$ and $p_2 = (3x^2 - 1)/2$. Consecutive polynomials satisfy

$$p_{n+1}(x) = \frac{2n+1}{n+1} x p_n(x) - \frac{n}{n+1} p_{n-1}(x). \quad (3.138)$$

3.5.2 Chebyshev polynomials

In the plane the angular factor is elementary. The real part $\operatorname{Re}(x + iy)^n = r^n \cos(n\theta)$ is harmonic, being the real part of an analytic function (Section 2.6), and $\cos(n\theta)$ is a polynomial in $x = \cos \theta$. That polynomial is the Chebyshev polynomial

$$T_n(x) = \cos(n \arccos x). \quad (3.139)$$

Chebyshev's equation $(1 - x^2)y'' - xy' + n^2y = 0$ imposes the same two-step relation on a power series,

$$(k + 2)(k + 1)a_{k+2} = (k^2 - n^2)a_k, \quad (3.140)$$

and the chain terminates at degree n . The addition formula $\cos((n \pm 1)\theta) = \cos(n\theta)\cos\theta \mp \sin(n\theta)\sin\theta$ adds to the recurrence

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n \geq 1, \quad (3.141)$$

starting from $T_0 = 1$ and $T_1 = x$. Hence $T_2 = 2x^2 - 1$ and $T_3 = 4x^3 - 3x$.

3.5.3 Spherical harmonics and Laguerre polynomials

The hydrogen atom adds the Coulomb potential to the Laplacian. In the units $\hbar = m = e = 1$,

$$-\frac{1}{2}\Delta\psi - \frac{1}{r}\psi = E\psi. \quad (3.142)$$

Spherical coordinates separate a solution as $R(r)Y_{lm}(\theta, \phi)$. The angular factor is a spherical harmonic, and its polar part is built from the same Legendre polynomial: the magnetic index $m = 0$ gives $p_l(\cos\theta)$, while $m \neq 0$ differentiates that polynomial and multiplies by a power of $\sin\theta$,

$$Y_{lm}(\theta, \phi) \propto e^{im\phi} (1 - \mu^2)^{|m|/2} \frac{d^{|m|} p_l}{d\mu^{|m|}}(\mu), \quad \mu = \cos\theta. \quad (3.143)$$

The bound energies are $E = -1/(2n^2)$ for integers $n \geq 1$ and orbital indices $0 \leq l \leq n - 1$. The radial factor is an associated Laguerre polynomial,

$$R_{nl}(r) \propto e^{-\rho/2} \rho^l L_{n-l-1}^{(2l+1)}(\rho), \quad \rho = \frac{2r}{n}. \quad (3.144)$$

The polynomial $L_n^{(\alpha)}$ solves $xy'' + (\alpha + 1 - x)y' + ny = 0$, so the radial factor has upper index $\alpha = 2l + 1$. The case $\alpha = 0$ is Laguerre's equation $xy'' + (1 - x)y' + ny = 0$. The origin is a regular singular point. The Frobenius series $y = \sum_{k \geq 0} a_k x^k$ has indicial root 0, and the coefficients satisfy

$$(k + 1)^2 a_{k+1} = (k - n) a_k. \quad (3.145)$$

For n a nonnegative integer the series stops at degree n . Rodrigues' formula is

$$L_n(x) = \frac{e^x}{n!} \frac{d^n}{dx^n} (x^n e^{-x}), \quad (3.146)$$

so $L_0 = 1$ and $L_1 = 1 - x$. The recurrence is

$$L_{n+1}(x) = \frac{2n + 1 - x}{n + 1} L_n(x) - \frac{n}{n + 1} L_{n-1}(x). \quad (3.147)$$

3.5.4 Hermite polynomials

The harmonic oscillator is the quadratic potential. In the units $\hbar = m = \omega = 1$,

$$-\frac{1}{2}\psi'' + \frac{1}{2}x^2\psi = E\psi. \quad (3.148)$$

The ground-state factor $e^{-x^2/2}$ removes the quadratic term: substituting $\psi = e^{-x^2/2}y$ cancels $x^2\psi$ against the second derivative of the Gaussian and leaves Hermite's equation

$$y'' - 2xy' + \lambda y = 0, \quad \lambda = 2E - 1. \quad (3.149)$$

The power series $y = \sum_{k \geq 0} a_k x^k$ then has

$$(k+2)(k+1)a_{k+2} = (2k - \lambda)a_k. \quad (3.150)$$

The choice $\lambda = 2n$ for a nonnegative integer n stops the chain at degree n , and the energy is $E = n + \frac{1}{2}$. Rodrigues' formula for the resulting polynomial is

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} (e^{-x^2}), \quad (3.151)$$

so $H_0 = 1$, $H_1 = 2x$, and $H_2 = 4x^2 - 2$. Consecutive polynomials satisfy

$$H_{n+1}(x) = 2x H_n(x) - 2n H_{n-1}(x). \quad (3.152)$$

3.5.5 Bessel functions

A circular membrane, or a wave of fixed frequency in the plane, separates in polar coordinates. Solutions $R(r)e^{in\theta}$ of the Helmholtz equation $\Delta u + \kappa^2 u = 0$ satisfy

$$r^2 R'' + rR' + (\kappa^2 r^2 - n^2)R = 0. \quad (3.153)$$

The change of scale $x = \kappa r$ puts this in the standard form $x^2 y'' + xy' + (x^2 - \nu^2)y = 0$ with $\nu = n$. The point $x = 0$ is a regular singular point: the indicial equation is $r^2 - \nu^2 = 0$. The solution that stays bounded at the origin is the Bessel function of the first kind,

$$J_\nu(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m! \Gamma(m + \nu + 1)} \left(\frac{x}{2}\right)^{2m+\nu}. \quad (3.154)$$

For $\nu = 0$ one has $\Gamma(m+1) = m!$, so $J_0(x) = \sum_{m \geq 0} (-1)^m (x/2)^{2m} / (m!)^2$. On an interval the series solution was a sine. On a disk it is $J_n(\kappa r)$ times an angular factor, and there is no discrete sine transform of that radial profile. Section 5.4 takes up the weighted integrals of p_n , T_n , L_n , and H_n ; those integrals are the numerical-analysis use of the functions defined here.

3.6 Nonlinear equations

The expansions of the previous sections use linearity twice: a sum of solutions is a solution, and the coefficients are inner products against a fixed family. Both fail when the unknown multiplies itself. This section records the model cases in which an explicit description still exists [30]. The lift that returns such an equation to a linear evolution in a larger space is Chapter 15.

3.6.1 Shocks

The model first-order nonlinear equation is

$$u_t + a(u) u_x = 0, \quad u(x, 0) = \phi(x). \quad (3.155)$$

Along a curve satisfying $dx/dt = a(u)$, the chain rule gives

$$\frac{d}{dt} u(x(t), t) = u_t + u_x \frac{dx}{dt} = u_t + a(u) u_x = 0. \quad (3.156)$$

Thus u is constant on the curve, equal to its value $\phi(x(0))$ at $t = 0$. The slope $a(u)$ is therefore constant as well, and the characteristic through ξ is the straight line

$$x = \xi + t a(\phi(\xi)), \quad u(x, t) = \phi(\xi). \quad (3.157)$$

For Burgers' equation (2.15) one has $a(u) = u$. The initial datum $\phi(x) = x$ produces the lines $x = \xi(1 + t)$, hence $\xi = x/(1 + t)$ and

$$u(x, t) = \frac{x}{1 + t}, \quad t > -1. \quad (3.158)$$

The characteristics spread, and the solution stays smooth. The datum $\phi(x) = 1 - x$ of Figure 2.3 does the opposite: the line through ξ is $x = \xi + t(1 - \xi)$, so for $t < 1$

$$u(x, t) = \frac{1 - x}{1 - t}. \quad (3.159)$$

As $t \rightarrow 1^-$ the denominator vanishes and every such line meets the point $(1, 1)$. Past that time the classical solution cannot be continued as a single-valued function.

The continuation is a weak solution of the conservation law. Write $a = A'$, so that (3.155) is $u_t + \partial_x A(u) = 0$. For Burgers, $A(u) = u^2/2$. For every test function ψ that vanishes for large $|x|$ and large t ,

$$\int_0^\infty \int_{\mathbb{R}} (u \psi_t + A(u) \psi_x) dx dt + \int_{\mathbb{R}} \phi(x) \psi(x, 0) dx = 0. \quad (3.160)$$

The identity is integration by parts on $u_t + A(u)_x = 0$: the volume terms cancel, and the initial trace of ψ produces the second integral. A function with a jump across a curve $x = \zeta(t)$ can satisfy (3.160) without satisfying the differential equation on that curve.

The jump is fixed by the Rankine–Hugoniot relation. Let $s = \zeta'$, and write u_l and u_r for the limits from the left and from the right of ζ . Integrate $u_t + \partial_x A(u) = 0$ from $\zeta(t) - \varepsilon$ to $\zeta(t) + \varepsilon$ and apply the Leibniz rule to the moving limits. The integral of the bounded function u over an interval of length 2ε tends to 0 with ε , and the remaining terms give

$$s = \frac{A(u_r) - A(u_l)}{u_r - u_l}, \quad (3.161)$$

provided $u_r \neq u_l$. For Burgers this is $s = (u_l + u_r)/2$.

The differential equation and the jump condition together do not select a unique continuation. The admissible shock is the one that characteristics run into. If a is increasing, that entropy condition reads

$$a(u_l) > s > a(u_r). \quad (3.162)$$

The wave behind the discontinuity is faster than the wave ahead of it, so the front is overtaken. If a is decreasing, the inequalities reverse. A jump that violates (3.162) is a weak solution of (3.160) and is discarded on these grounds.

An example shows why the condition is needed. Take Burgers' equation with $\phi(x) = 0$ for $x < 0$ and $\phi(x) = 1$ for $x > 0$. Characteristics from the left have speed 0 and those from the right have speed 1, so no characteristic reaches the wedge $0 < x < t$, and two functions fill it. The first is a single jump, $u = 0$ for $x < t/2$ and $u = 1$ for $x > t/2$. By (3.161) the speed is $s = (0 + 1)/2 = \frac{1}{2}$, so it is a weak solution. The second is the rarefaction wave

$$u(x, t) = \begin{cases} 0, & x \leq 0, \\ x/t, & 0 < x < t, \\ 1, & x \geq t. \end{cases} \quad (3.163)$$

It is continuous, and on the wedge $u_t + uu_x = -x/t^2 + (x/t)(1/t) = 0$, so it is a weak solution as well. Each ray $x/t = \text{const}$ is a characteristic leaving the origin. The jump violates (3.162), because $a(u_l) = 0$ is not greater than $s = \frac{1}{2}$: the characteristics leave the discontinuity on both sides instead of running into it, and any small perturbation of the data unfolds it into the fan. The rarefaction wave is the admissible solution. For the reversed data $u_l = 1$, $u_r = 0$ the speed is again $s = \frac{1}{2}$, and $1 > \frac{1}{2} > 0$ holds, so the jump is admissible: characteristics converge on it from both sides, and no fan can replace it because characteristics would cross. For a convex flux the rule is therefore simple. If $u_l < u_r$ the characteristics diverge and the solution is continuous, $u = (A')^{-1}(x/t)$ in the fan. If $u_l > u_r$ they converge and the solution is a shock. The same selection is made by adding a small viscosity: the solutions of $u_t + uu_x = \nu u_{xx}$ converge, as $\nu \rightarrow 0$, to the entropy solution.

3.6.2 The KdV soliton

Dispersion can balance the steepening that produced (3.159). The Korteweg–de Vries equation

$$u_t + 6u u_x + u_{xxx} = 0 \quad (3.164)$$

has, for solutions that decay rapidly at infinity, the conserved integrals

$$\int_{\mathbb{R}} u \, dx, \quad \int_{\mathbb{R}} u^2 \, dx, \quad \int_{\mathbb{R}} \left(\frac{1}{2} (u_x)^2 - u^3 \right) dx. \quad (3.165)$$

The first is the integral of (3.164): $\int u_{xxx} = 0$ and $\int uu_x = \frac{1}{2} \int (u^2)_x = 0$. The second follows on multiplying by $2u$ and integrating by parts, using $\int u u_{xxx} = -\int u_x u_{xx} = 0$ and $\int u^2 u_x = \frac{1}{3} \int (u^3)_x = 0$. The third has the same structure. Integrating $u_x u_{xt}$ by parts,

$$\frac{d}{dt} \int \left(\frac{1}{2} u_x^2 - u^3 \right) dx = \int (-u_{xx} - 3u^2) u_t \, dx, \quad (3.166)$$

and (3.164) says $u_t = -(u_{xx} + 3u^2)_x$, because $(3u^2)_x = 6u u_x$. With $m = u_{xx} + 3u^2$ the right side is $\int m m_x \, dx = \frac{1}{2} [m^2]_{-\infty}^{\infty} = 0$.

A traveling wave $u(x, t) = f(x - ct)$ reduces (3.164) to $-cf' + f''' + 6ff' = 0$. One integration gives $-cf + f'' + 3f^2 = a$. Multiplication by f' and a second integration give

$$-cf^2 + (f')^2 + 2f^3 = 2af + b. \quad (3.167)$$

The localized wave corresponds to $a = b = 0$, so $(f')^2 = f^2(c - 2f)$. Set $g^2 = c - 2f$ with $c > 0$ and $f < c/2$. Then $f' = -gg'$ and $2g' = g^2 - c$. The solution $g = -\sqrt{c} \tanh(\sqrt{c}\xi/2)$, $\xi = x - ct$, satisfies that first-order equation, and $f = (c - g^2)/2$ becomes

$$f(\xi) = \frac{c}{2} \operatorname{sech}^2\left(\frac{\sqrt{c}}{2}\xi\right). \quad (3.168)$$

The profile propagates at speed c without changing shape. The height $c/2$ and the width $1/\sqrt{c}$ are tied to the same constant: a taller wave is narrower and faster. This is the balance between $6uu_x$ and u_{xxx} that the inviscid Burgers equation does not have.

3.6.3 Fluid equations and acoustics

The same conservation structure, written for a density and a velocity, is the starting point of the fluid equations. Conservation of mass on an arbitrary region D says that the rate of change of $\int_D \rho$ equals the inward flux $-\int_{\partial D} \rho v \cdot \nu \, dS$. The divergence theorem converts this into

$$\rho_t + \nabla \cdot (\rho v) = 0. \quad (3.169)$$

Conservation of momentum, with an external force set to zero and with a pressure p , likewise gives

$$(\rho v)_t + \nabla \cdot (\rho v \otimes v) + \nabla p = 0. \quad (3.170)$$

Expanding the derivatives and cancelling $\rho_t + \nabla \cdot (\rho v)$ against (3.169) produces

$$v_t + (v \cdot \nabla)v + \frac{1}{\rho} \nabla p = 0. \quad (3.171)$$

If $p = p(\rho)$ and the density is a positive constant, $\nabla \cdot v = 0$ and (3.171) is Euler's equation. Section 3.6.4 restores the viscous stress.

Sound is the linearization of (3.169) and (3.171) about a rest state. Set $\rho = \rho_0 + \varepsilon \rho_1$ and $v = \varepsilon v_1$, with $\rho_0 > 0$ constant, and let $p = f(\rho)$. Then $p = f(\rho_0) + f'(\rho_0)(\rho - \rho_0)$ up to order ε^2 . Substituting and keeping the order ε terms,

$$\partial_t \rho_1 + \rho_0 \nabla \cdot v_1 = 0, \quad \partial_t v_1 + \frac{f'(\rho_0)}{\rho_0} \nabla \rho_1 = 0. \quad (3.172)$$

Differentiate the first equation in time and substitute $\partial_t v_1$ from the second:

$$\partial_{tt} \rho_1 - f'(\rho_0) \Delta \rho_1 = 0. \quad (3.173)$$

If the initial velocity is irrotational, the velocity remains irrotational and satisfies the same wave equation. The propagation speed is $\sqrt{f'(\rho_0)}$, which is the constant c of Section 3.4. The kernels and series of that section are the linear acoustic approximation of (3.169)–(3.171).

3.6.4 The Navier–Stokes equations

Air past a wing and the flow of water are described by restoring viscosity to (3.171). For a constant density the velocity and the pressure satisfy the incompressible Navier–Stokes equations

$$v_t + (v \cdot \nabla)v = -\frac{1}{\rho} \nabla p + \nu \Delta v, \quad \nabla \cdot v = 0, \quad (3.174)$$

with kinematic viscosity $\nu > 0$. The new term is the one absent from Euler's equation. If the density varies, (3.169) is kept and the viscous contribution to (3.170) is the divergence of the stress $\tau = \mu(\nabla v + (\nabla v)^\top) + \lambda(\nabla \cdot v)I$, where $\mu = \rho\nu$ is the shear viscosity and λ is a second viscosity coefficient. For constant μ and for $\nabla \cdot v = 0$ that divergence equals $\mu\Delta v$, which becomes $\nu\Delta v$ after division by ρ . On a wall at rest the fluid does not slip, so $v = 0$ there. The viscous Burgers equation of Section 2.2.4 is the scalar model of the same balance, with νu_{xx} in the role of $\nu\Delta v$.

The form of the viscous stress follows from three assumptions. The force across a surface element with unit normal ν is $\tau\nu$, where τ is a stress tensor. The tensor depends linearly on the velocity gradient ∇v , which defines a Newtonian fluid. A rigid rotation $v = \omega \times x$ exerts no stress. The gradient of a rigid rotation is antisymmetric, so only the symmetric part $D = \frac{1}{2}(\nabla v + (\nabla v)^\top)$ can enter. An isotropic linear map from symmetric matrices to symmetric matrices has the form $D \mapsto 2\mu D + \lambda(\text{tr } D)I$, and $\text{tr } D = \nabla \cdot v$. This is the stress τ above. The force on a fluid region is $\int_{\partial D} \tau\nu \, dS$, which the divergence theorem turns into $\int_D \nabla \cdot \tau$, with

$$(\nabla \cdot \tau)_i = \sum_j \partial_j \tau_{ij} = \mu \Delta v_i + (\mu + \lambda) \partial_i (\nabla \cdot v). \quad (3.175)$$

For $\nabla \cdot v = 0$ only $\mu\Delta v_i$ remains, which is the viscous term of (3.174).

The pressure in (3.174) is the field that enforces the constraint $\nabla \cdot v = 0$ at every time. Take the divergence of (3.174). The divergence of v_t and of Δv vanishes, and

$$\Delta p = -\rho \sum_{i,j} \partial_i v_j \partial_j v_i. \quad (3.176)$$

Thus every instant of the evolution contains a Poisson problem for p , of the kind represented by (3.68), with a right-hand side determined by the velocity.

Viscosity dissipates kinetic energy. Multiply (3.174) by v and integrate over a bounded domain on which $v = 0$, or over $\mathbb{R}^3$ for a velocity that decays together with its derivatives. The time term is $\frac{1}{2} \frac{d}{dt} \int |v|^2$. The convective term is $\frac{1}{2} \int v \cdot \nabla(|v|^2) = -\frac{1}{2} \int |v|^2 \nabla \cdot v$, which vanishes because $\nabla \cdot v = 0$, and the boundary integral produced by the divergence theorem vanishes because $v = 0$ there. The pressure term $-\int v \cdot \nabla p$ vanishes for the same two reasons. The viscous term integrates by parts to $-\nu \int |\nabla v|^2$, where $|\nabla v|^2 = \sum_{i,j} (\partial_i v_j)^2$. Therefore

$$\frac{d}{dt} \frac{1}{2} \int |v|^2 \, dx = -\nu \int |\nabla v|^2 \, dx \leq 0. \quad (3.177)$$

The integral of $|v|^2$ is nonincreasing, and $\nu = 0$ returns the conservation of kinetic energy in Euler's equation. The space-time integral of $|\nabla v|^2$ is controlled by the initial energy.

The size of the nonlinear term relative to viscosity is the Reynolds number. Let L be a length and U a speed, and introduce $x = Lx'$, $t = (L/U)t'$, $v = Uv'$, and $p = \rho U^2 p'$. Substitution into (3.174) and division by U^2/L produce

$$\partial_t v' + (v' \cdot \nabla') v' = -\nabla' p' + \frac{1}{\text{Re}} \Delta' v', \quad \text{Re} = \frac{UL}{\nu}, \quad (3.178)$$

together with $\nabla' \cdot v' = 0$. A large value of Re makes the viscous term small on the scale of the domain, and the balance approaches Euler's equation. A small value makes the linear Stokes balance $-\nu\Delta v + \frac{1}{\rho}\nabla p = 0$, still constrained by $\nabla \cdot v = 0$, the leading description; the convective term is then a perturbation. Chapter 15 treats the lift of a dissipative nonlinearity in the regime where viscosity dominates that term [37, 17].

The energy identity (3.177) bounds $\int |v|^2$ for all later times. In two space dimensions this bound is enough to continue a smooth divergence-free solution for all positive times. In three dimensions the same bound does not control $(v \cdot \nabla)v$, and it is open whether every smooth, finite-energy, divergence-free initial velocity remains smooth, or whether the velocity of some solution becomes singular in finite time. That question is one of the Millennium Prize problems.

Chapter 4

The Cauchy theory

Section 2.6 showed that a complex derivative forces the Cauchy–Riemann equations, and that each component of a holomorphic function is harmonic. The integral theory that upgrades that derivative is Cauchy’s theorem. A single complex derivative, with no continuity assumed of the derivative, already implies that integrals over triangles vanish, that a primitive exists on a disc, that the values inside a circle are the boundary integral against $1/(\zeta - z)$, and that the function is a power series. Derivatives of every order exist, and a uniform limit of holomorphic functions is holomorphic. The development below follows the route from Goursat’s theorem through the keyhole contour, the Cauchy integral formulas, and the consequences collected by Stein and Shakarchi [46]. The last two sections read those formulas as the plane potential theory of Chapter 3: the logarithm, the Poisson kernel, and conformal maps.

4.1 Line integrals

Let $\gamma: [a, b] \rightarrow \mathbb{C}$ be piecewise continuously differentiable and let f be continuous on the image of γ . The integral of f along γ is

$$\int_{\gamma} f(z) \, dz = \int_a^b f(\gamma(t)) \gamma'(t) \, dt. \quad (4.1)$$

An orientation-preserving reparametrization leaves the value unchanged, because the chain rule inserts a derivative that the substitution rule cancels. Reversing the endpoints changes the sign. If f has a primitive on an open set containing the image, $F' = f$, the chain rule gives $\frac{d}{dt} F(\gamma(t)) = f(\gamma(t))\gamma'(t)$, so the integral equals $F(\gamma(b)) - F(\gamma(a))$. The integral of a derivative over a closed path is therefore zero. The functions 1 and $z - z_0$ have primitives z and $(z - z_0)^2/2$.

4.2 Goursat’s theorem

The basic vanishing uses no continuity of the derivative. Let T be a closed triangle, including its interior, contained in an open set where f' exists at every point. Then

$$\int_{\partial T} f(z) \, dz = 0. \quad (4.2)$$

Join the midpoints of the sides. The triangle splits into four congruent triangles of half the diameter and half the perimeter. Orient each of them consistently with the outer boundary. The integrals over the three internal edges cancel in pairs, so the integral over ∂T is the sum of the four

smaller integrals. At least one of the four, called T_1 , carries at least a quarter of the absolute value. Repeat. Figure 4.1 shows the first bisection and the diagonal that splits a rectangle. If the original perimeter is L and the original diameter is D , the n th triangle T_n satisfies

$$\left| \int_{\partial T} f \right| \leq 4^n \left| \int_{\partial T_n} f \right|, \quad \text{diam}(T_n) = 2^{-n}D, \quad \text{length}(\partial T_n) = 2^{-n}L. \quad (4.3)$$

The solid triangles are nested, closed, and bounded, and their diameters tend to zero. A nested sequence of that kind in the plane has exactly one point in the intersection: this is the nested-interval form of completeness of $\mathbb{R}$, applied to the real and imaginary parts. Call the point z_0 .

Complex differentiability at z_0 means

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + \psi(z)(z - z_0), \quad (4.4)$$

with $\psi(z) \rightarrow 0$ as $z \rightarrow z_0$. The constant and the linear term integrate to zero over ∂T_n , because each has a primitive. The integral of f over ∂T_n therefore equals the integral of $\psi(z)(z - z_0)$. On T_n one has $|z - z_0| \leq 2^{-n}D$. If $\varepsilon_n = \sup_{T_n} |\psi|$, then

$$\left| \int_{\partial T_n} f \right| \leq \varepsilon_n 2^{-n}D \cdot 2^{-n}L, \quad (4.5)$$

and the comparison of the four-way split gives $|\int_{\partial T} f| \leq \varepsilon_n DL$. Since $\varepsilon_n \rightarrow 0$, the integral over the original triangle is zero.

A rectangle is two triangles. Draw one diagonal and orient both triangles so that the diagonal is traversed once in each direction. The diagonal cancels, and the sum of the two boundary integrals is the integral over the boundary of the rectangle. Each triangle contributes zero, so

$$\int_{\partial R} f(z) dz = 0 \quad (4.6)$$

whenever f is holomorphic on an open set containing the rectangle R and its interior.

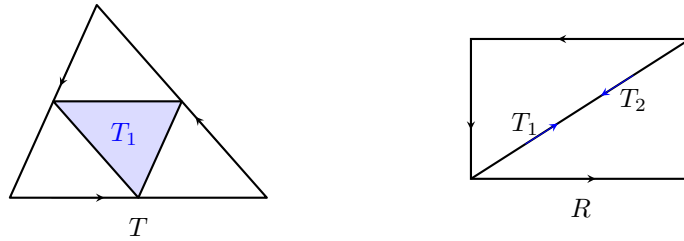


Figure 4.1: Left: the first bisection in the proof of (4.2). The three midlines cancel in pairs, and T_1 is a triangle whose integral is at least a quarter of the original. Right: a rectangle is the union of two triangles. The diagonal is traversed once in each direction.

4.3 Primitives in a disc

A holomorphic function on an open disc has a primitive there. After a translation, take the disc centered at the origin. For $z = x + iy$ in the disc let γ_z be the polygonal path that runs horizontally from 0 to $\tilde{z} = x$ and then vertically from $\tilde{z}$ to z , as in Figure 4.2. Define

$$F(z) = \int_{\gamma_z} f(w) dw. \quad (4.7)$$

The path is unambiguous, so F is well defined before any path-independence has been proved. The claim is $F' = f$.

It is enough to treat a horizontal increment and a vertical increment, and then to combine them. Let $h \in \mathbb{R}$ be small enough that the rectangle with corners z , $z + h$, $x + h$, and x lies in the disc. The path γ_{z+h} runs from 0 to $x + h$ and then up to $z + h$. Subtracting γ_z cancels the segment from 0 to x . What remains is the integral from x to $x + h$, up to $z + h$, minus the integral up from x to z . The boundary of the rectangle with those corners has integral zero by (4.6). The two vertical sides and the bottom account for exactly that difference, so the horizontal side from z to $z + h$ is what survives:

$$F(z + h) - F(z) = \int_{[z, z+h]} f(w) dw. \quad (4.8)$$

The same identity holds for a purely imaginary increment, with the rectangle standing on a vertical side. A general small increment $h = h_1 + ih_2$ is the sum of the two, so $F(z + h) - F(z)$ equals the integral of f along the horizontal segment from z to $z + h_1$ followed by the vertical segment from $z + h_1$ to $z + h$. Those two segments and the straight segment from $z + h$ back to z form a triangle, the second panel of Figure 4.2. Its integral vanishes by (4.2), and therefore $F(z + h) - F(z)$ equals the integral of f along the straight segment η from z to $z + h$.

Continuity of f , which follows from differentiability, gives $f(w) = f(z) + \psi(w)$ with $\psi(w) \rightarrow 0$ as $w \rightarrow z$. Then

$$F(z + h) - F(z) = f(z) \int_{\eta} dw + \int_{\eta} \psi(w) dw = f(z) h + \int_{\eta} \psi. \quad (4.9)$$

The constant 1 has primitive w , so the first integral equals h . The second is at most $\sup_{\eta} |\psi| |h|$. Dividing by h and sending $h \rightarrow 0$ yields $F'(z) = f(z)$.

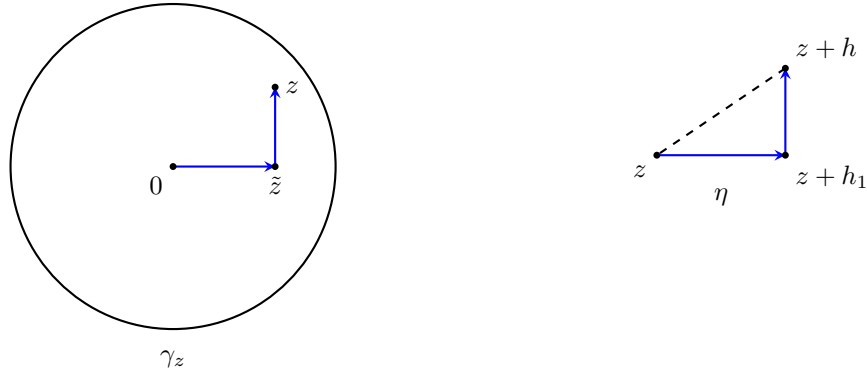


Figure 4.2: Left: the path γ_z that defines the primitive (4.7). Right: a general increment decomposes into a horizontal segment and a vertical segment. With the dashed side they bound a triangle, so the broken path equals the straight segment η .

Every closed piecewise smooth path in the disc now integrates f to zero, because the integral equals the change in F between two visits to the same endpoint. In particular, if f is holomorphic on an open set containing a circle and the disc it bounds, the integral of f over the circle vanishes: the circle sits in a slightly larger disc on which the primitive exists.

The same construction applies to any closed curve for which the interior is plain and a neighborhood of the curve and its interior can be travelled by horizontal and vertical segments. Call such a curve a toy contour, and orient it so that the interior lies to the left. Circles, triangles, and rectangles are toy contours. So is a keyhole: two nearly complete circles, one large and one small,

joined by a narrow corridor. On a keyhole the integral of a holomorphic function is independent of the choice of horizontal-and-vertical path used to define a primitive in the interior, because the difference of two such paths is a chain of rectangles to which (4.6) applies. The primitive then forces the integral over the keyhole itself to vanish. The same vanishing holds for a semicircle closed by a diameter, for a sector, and for a rectangle with a small semicircular indentation.

4.4 Two real integrals

Cauchy's theorem evaluates real integrals by moving the path. The first example is the Fourier transform of a Gaussian. The integral $\int_{-\infty}^{\infty} e^{-\pi x^2} dx = 1$ is the normalization used for the heat kernel in Section 3.1. For $\xi > 0$ the same Gaussian, shifted into the complex plane, gives

$$e^{-\pi \xi^2} = \int_{-\infty}^{\infty} e^{-\pi x^2} e^{-2\pi i x \xi} dx. \quad (4.10)$$

The function $f(z) = e^{-\pi z^2}$ is entire. Integrate it over the rectangle with vertices $-R$, R , $R + i\xi$, and $-R + i\xi$, oriented counterclockwise, as in the left panel of Figure 4.3. The integral is zero. On the bottom side it is $\int_{-R}^R e^{-\pi x^2} dx$, which tends to 1. On the right side $z = R + iy$ with $0 \leq y \leq \xi$, and

$$(R + iy)^2 = R^2 - y^2 + 2iRy, \quad |e^{-\pi(R+iy)^2}| = e^{-\pi R^2 + \pi y^2} \leq e^{-\pi R^2 + \pi \xi^2}. \quad (4.11)$$

The length of the side is ξ , so the integral is $O(e^{-\pi R^2})$ and tends to zero. The left side does the same. On the top side $z = x + i\xi$ runs from R down to $-R$, and

$$(x + i\xi)^2 = x^2 - \xi^2 + 2ix\xi, \quad (4.12)$$

so the integral equals $-e^{\pi \xi^2} \int_{-R}^R e^{-\pi x^2} e^{-2\pi i x \xi} dx$. Sending $R \rightarrow \infty$ leaves $0 = 1 - e^{\pi \xi^2}$ times the integral in (4.10). For $\xi < 0$ the rectangle lies in the lower half-plane.

The second example indents the contour around a singularity on the path of integration:

$$\int_0^{\infty} \frac{1 - \cos x}{x^2} dx = \frac{\pi}{2}. \quad (4.13)$$

Consider $f(z) = (1 - e^{iz})/z^2$ on the contour in the right panel of Figure 4.3: the real axis from $-R$ to $-\varepsilon$, the semicircle of radius ε clockwise, the real axis from ε to R , and the semicircle of radius R counterclockwise. The function is holomorphic inside this contour, because the origin has been left outside, so the integral vanishes. On the large semicircle $|e^{iz}| = e^{-R \sin \theta} \leq 1$, hence $|1 - e^{iz}| \leq 2$. The length is πR and the integrand is $O(1/R^2)$, so the integral is $O(1/R)$ and tends to zero. Near the origin, $e^{iz} = 1 + iz + O(z^2)$, and therefore

$$f(z) = -\frac{i}{z} + E(z) \quad (4.14)$$

with E bounded as $z \rightarrow 0$. On the small semicircle $z = \varepsilon e^{i\theta}$ with θ running from π down to 0, one has $dz = i\varepsilon e^{i\theta} d\theta$, and

$$\int \frac{-i}{z} dz = \int_{\pi}^0 (-i) \cdot i d\theta = -\pi. \quad (4.15)$$

The integral of the bounded remainder is $O(\varepsilon)$ and tends to zero. The vanishing of the whole contour therefore yields

$$\int_{-\infty}^{\infty} \frac{1 - e^{ix}}{x^2} dx = \pi. \quad (4.16)$$

The imaginary part $(-\sin x)/x^2$ is odd and integrates to zero. The real part is even, and half of it is (4.13).

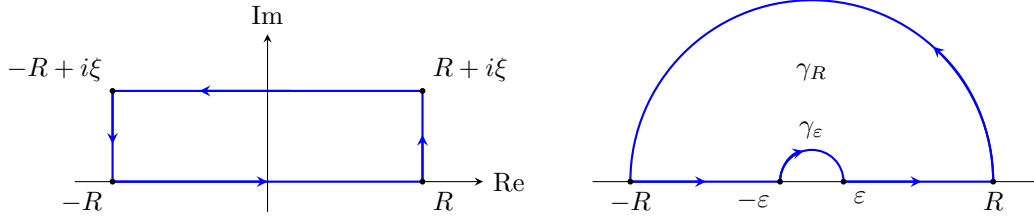


Figure 4.3: Contours for the two real integrals. Left: the rectangle that shifts $e^{-\pi z^2}$ by $i\xi$. The vertical sides vanish as $R \rightarrow \infty$, and the top side carries the Fourier factor $e^{-2\pi i x \xi}$. Right: the indented semicircle for $(1 - e^{iz})/z^2$. The small semicircle is clockwise, so the origin stays outside the contour.

4.5 Cauchy's integral formulas

Suppose f is holomorphic on an open set containing a closed disc, and let C be the boundary circle with positive orientation. For every z inside the disc,

$$f(z) = \frac{1}{2\pi i} \int_C \frac{f(\zeta)}{\zeta - z} d\zeta. \quad (4.17)$$

Fix z inside the disc and form a keyhole $\Gamma_{\delta, \varepsilon}$, drawn in Figure 4.4, that follows C except for a corridor of width δ leading to a circle of radius ε about z . The corridor is oriented so that the interior of the keyhole, which excludes z , lies to the left. The quotient $F(\zeta) = f(\zeta)/(\zeta - z)$ is holomorphic on a neighborhood of that keyhole and its interior, so its integral over the keyhole vanishes. Let $\delta \rightarrow 0$. The two sides of the corridor are the same segment in opposite directions, and continuity of F makes their integrals cancel. What remains is the integral of F over C , oriented positively, plus the integral over the small circle C_ε , oriented clockwise. Write

$$F(\zeta) = \frac{f(\zeta) - f(z)}{\zeta - z} + \frac{f(z)}{\zeta - z}. \quad (4.18)$$

Holomorphy at z bounds the first quotient. Its integral over C_ε is at most a constant times the length $2\pi\varepsilon$, and tends to zero with ε . On C_ε one has $\zeta = z + \varepsilon e^{-it}$ as t runs from 0 to 2π , the minus sign recording the clockwise orientation, and $d\zeta = -i\varepsilon e^{-it} dt$. Thus

$$\int_{C_\varepsilon} \frac{f(z)}{\zeta - z} d\zeta = f(z) \int_0^{2\pi} (-i) dt = -2\pi i f(z). \quad (4.19)$$

The keyhole identity in the limit $\varepsilon \rightarrow 0$ is therefore the integral of F over C , minus $2\pi i f(z)$, equal to zero. This is (4.17).

The same argument with a rectangular keyhole gives the formula when C is replaced by the boundary of a rectangle. If z lies outside the contour, F itself is holomorphic inside and the integral is zero.

Differentiating under the integral produces every higher derivative. If f is holomorphic on an open set Ω , then f has complex derivatives of every order on Ω . If $C \subset \Omega$ is a circle whose interior also lies in Ω , then for every z inside C and every $n \geq 0$,

$$f^{(n)}(z) = \frac{n!}{2\pi i} \int_C \frac{f(\zeta)}{(\zeta - z)^{n+1}} d\zeta. \quad (4.20)$$

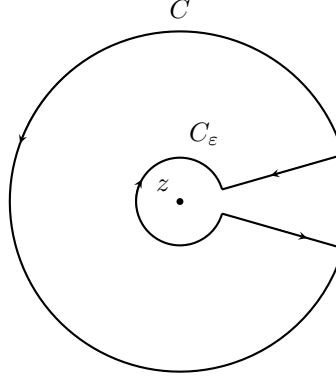


Figure 4.4: The keyhole $\Gamma_{\delta,\varepsilon}$ in the proof of (4.17). The outer circle is counterclockwise and the inner circle, centered at the interior point z , is clockwise. The two sides of the corridor cancel as its width tends to zero, and z remains outside the contour.

The case $n = 0$ is (4.17). Assume the formula for order $n - 1$. The difference quotient of $f^{(n-1)}$ is

$$\frac{f^{(n-1)}(z+h) - f^{(n-1)}(z)}{h} = \frac{(n-1)!}{2\pi i} \int_C f(\zeta) \frac{1}{h} \left(\frac{1}{(\zeta - z - h)^n} - \frac{1}{(\zeta - z)^n} \right) d\zeta. \quad (4.21)$$

The algebraic identity $A^n - B^n = (A - B) \sum_{j=0}^{n-1} A^{n-1-j} B^j$, with $A = 1/(\zeta - z - h)$ and $B = 1/(\zeta - z)$, converts the difference of powers into

$$\frac{1}{h} (A^n - B^n) = \frac{1}{(\zeta - z - h)(\zeta - z)} \sum_{j=0}^{n-1} A^{n-1-j} B^j. \quad (4.22)$$

For small h both z and $z+h$ stay at a positive distance from C . Sending $h \rightarrow 0$ replaces each factor by $1/(\zeta - z)$, and the sum contains n equal terms. The difference quotient tends to the right-hand side of (4.20).

If in addition $|f| \leq M$ on a circle of radius R about z_0 , the length $2\pi R$ and the denominator R^{n+1} give the Cauchy inequalities

$$|f^{(n)}(z_0)| \leq \frac{n! M}{R^n}. \quad (4.23)$$

The integral formula is also a power series. Chapter 1 of Stein and Shakarchi records that a power series is holomorphic inside its disc of convergence. The converse is the present formula. Let D be a disc centered at z_0 whose closure lies in the domain of f , and let C be its boundary. For $z \in D$ and $\zeta \in C$,

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - z_0} \frac{1}{1 - (z - z_0)/(\zeta - z_0)} = \sum_{n=0}^{\infty} \frac{(z - z_0)^n}{(\zeta - z_0)^{n+1}}, \quad (4.24)$$

and the geometric series converges uniformly on C because $|z - z_0|/|\zeta - z_0|$ is strictly less than 1. Substitute into (4.17) and integrate term by term:

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n, \quad a_n = \frac{1}{2\pi i} \int_C \frac{f(\zeta)}{(\zeta - z_0)^{n+1}} d\zeta = \frac{f^{(n)}(z_0)}{n!}. \quad (4.25)$$

The series converges on every disc about z_0 whose closure lies in the domain. An entire function, holomorphic on all of $\mathbb{C}$, therefore has a power series about 0 with infinite radius of convergence. Termwise differentiation of (4.25) is a second proof that f' exists and is itself given by a power series.

4.6 Liouville's theorem and the identity theorem

If f is entire and $|f| \leq M$, the Cauchy inequality for $n = 1$ gives $|f'(z_0)| \leq M/R$ for every R . Letting $R \rightarrow \infty$ forces $f' = 0$. On the connected set $\mathbb{C}$ a function with vanishing derivative is constant. This is Liouville's theorem.

It implies the fundamental theorem of algebra. Let $P(z) = a_n z^n + \cdots + a_0$ with $a_n \neq 0$ and $n \geq 1$. If P had no root, then $1/P$ would be entire. For large $|z|$,

$$\frac{P(z)}{z^n} = a_n + \frac{a_{n-1}}{z} + \cdots + \frac{a_0}{z^n}, \quad (4.26)$$

so $|P(z)| \geq (|a_n|/2)|z|^n$ outside some disc. Thus $1/P$ is bounded at infinity. On the closed disc it is continuous, hence bounded. Liouville's theorem makes $1/P$ constant, which a nonconstant polynomial is not. So P has a root w_1 . Expanding in powers of $z - w_1$ and using $P(w_1) = 0$ factors $P(z) = (z - w_1)Q(z)$ with $\deg Q = n - 1$. Induction produces n roots and the factorization $P(z) = a_n(z - w_1) \cdots (z - w_n)$.

Holomorphic functions are rigid. Suppose f is holomorphic on a connected open set Ω and the zeros of f accumulate at a point $z_0 \in \Omega$. The power series of f at z_0 has every coefficient zero. If the first nonzero coefficient were a_m , one could write $f(z) = a_m(z - z_0)^m(1 + g(z - z_0))$ with $g(0) = 0$, and $1 + g$ would stay away from zero on a small punctured disc, contradicting an accumulation of zeros. Thus f vanishes on a disc about z_0 . The set of points at which f vanishes on a whole neighborhood is open, and it is closed in Ω : a limit of such points is a zero, and the series argument spreads the vanishing to a neighborhood. A nonempty set that is both open and closed in a connected Ω is all of Ω . Therefore f vanishes identically. The same statement for $f - g$ says that two holomorphic functions on a connected open set that agree on a nonempty open subset, or on a sequence of distinct points with a limit point in the set, agree everywhere. An analytic continuation, a holomorphic function on a larger connected open set that agrees with f on the original domain, is unique.

4.7 Further consequences

Morera's theorem is the converse of the triangle vanishing, and it uses the regularity just proved. Let f be continuous on an open disc, and suppose the integral of f over every triangle in the disc vanishes. The polygonal construction of (4.7) identified $F(z + h) - F(z)$ with the integral along the straight segment by vanishing integrals over rectangles and one triangle. A rectangle is two triangles, so the hypothesis gives the same vanishing. The same estimate as before gives $F' = f$. The derivative F' of the holomorphic function F is holomorphic by (4.20). Hence f is holomorphic.

The five descriptions met so far are therefore equivalent on an open set. Existence of f' , vanishing of all triangle integrals, the Cauchy integral formula on every compactly contained disc, existence of a power series at every point, and the Cauchy–Riemann equations for continuously differentiable components: each implies the others. The Cauchy–Riemann direction was proved in Section 2.6 as a necessary condition, and the sufficient direction is the real chain rule. If u and v

are C^1 and satisfy (2.45), then with $h = h_1 + ih_2$,

$$f(z_0 + h) - f(z_0) = (u_x - iu_y)h + o(|h|), \quad (4.27)$$

so $f'(z_0) = u_x + iv_x$. The content of the circle is that a mere complex derivative is already a power series.

Uniform limits preserve holomorphy. Let f_n be holomorphic on an open set Ω and converge to f uniformly on every compact subset. Then f is holomorphic on Ω . Indeed, on any disc whose closure lies in Ω , and on any triangle in that disc, Goursat's theorem gives $\int f_n = 0$. Uniform convergence passes the limit inside the integral, so $\int f = 0$. Morera's theorem says that f is holomorphic on the disc. The derivatives converge as well. Let Ω_δ be the set of points of Ω whose disc of radius δ still lies in Ω . If F is holomorphic on Ω and $z \in \Omega_\delta$, the formula (4.20) on the circle of radius δ about z yields

$$\sup_{z \in \Omega_\delta} |F'(z)| \leq \frac{1}{\delta} \sup_{\Omega} |F|. \quad (4.28)$$

Apply this to $F = f_n - f$. Uniform convergence of f_n on a neighborhood of the closure of Ω_δ implies uniform convergence of f'_n on Ω_δ . The same bound with $(\zeta - z)^{-k-1}$ gives the result for every order of derivative. A real uniform limit of continuously differentiable functions need not be differentiable: Weierstrass's theorem approximates every continuous function on a compact interval by polynomials. The complex statement has no such gap.

An integral depending on a parameter is holomorphic under a joint continuity assumption. Let $F(z, s)$ be defined for z in an open set Ω and $s \in [0, 1]$, holomorphic in z for each s , and continuous on $\Omega \times [0, 1]$. Then

$$f(z) = \int_0^1 F(z, s) ds \quad (4.29)$$

is holomorphic on Ω . The Riemann sums $f_n(z) = n^{-1} \sum_{k=1}^n F(z, k/n)$ are holomorphic. On a disc D whose closure lies in Ω , the function F is uniformly continuous, so $|F(z, s_1) - F(z, s_2)|$ is small whenever $|s_1 - s_2|$ is small, independently of $z \in D$. For large n the sum f_n is therefore uniformly close to f on D . The uniform-limit theorem makes f holomorphic on D , hence on Ω .

The Schwarz reflection principle extends a holomorphic function across a real interval on which it is real. Let Ω be open and symmetric with respect to the real axis, as in Figure 4.5, let Ω^+ and Ω^- be its intersections with the upper and lower half-planes, and let I be the interior of $\Omega \cap \mathbb{R}$ relative to the line. Suppose f is holomorphic on Ω^+ , extends continuously to I , and is real-valued on I . Define $F = f$ on $\Omega^+ \cup I$, and

$$F(z) = \overline{f(\bar{z})} \quad \text{for } z \in \Omega^-. \quad (4.30)$$

On the lower half-plane this is holomorphic: the power series $f(w) = \sum a_n(w - w_0)^n$ about a point of Ω^+ becomes, after conjugation of the coefficients and of the variable, a power series for F about $\bar{w}_0$. On I one has $f(x) = \overline{f(x)}$, so F is continuous across I . It remains to check holomorphy at points of I . Let D be a disc centered on I and contained in Ω . For a triangle $T \subset D$ that misses I , the integral of F vanishes because F is holomorphic on each open half-disc. If T meets I , split T into triangles each of which has at most one edge on I . Raise that edge by a small height ε , as in the right panel of Figure 4.5. The raised triangle lies entirely in one half-disc, so its integral vanishes. Continuity of F lets $\varepsilon \rightarrow 0$ and preserves the vanishing. Morera's theorem says that F is holomorphic on D , hence on Ω .

The same reflection is the complex form of the method of images. If $f = u + iv$ is real on the real axis, then $u(x, -y) = u(x, y)$ and $v(x, -y) = -v(x, y)$. The odd extension of a harmonic

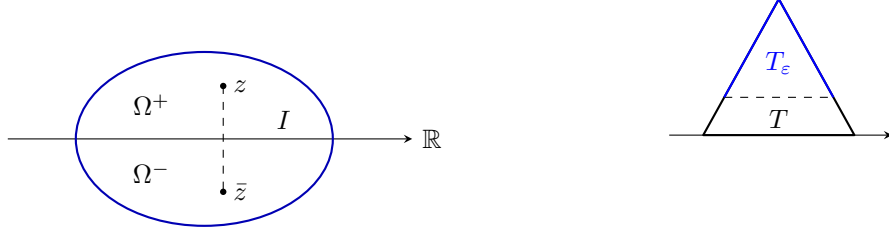


Figure 4.5: Left: a domain symmetric across the real axis. The value at $\bar{z}$ is the conjugate of the value at z when the boundary function on I is real. Right: a triangle with one side on I , and the raised triangle T_ϵ used in the Morera argument.

function that vanishes on the boundary line is the imaginary part of this construction, and it is the reflection that produced the half-space Green's function (3.77).

Runge's theorem is the complex form of Weierstrass approximation, with poles allowed where the complement requires them. A holomorphic function on a neighborhood of a compact set K can be approximated uniformly on K by rational functions whose poles lie in the complement of K . If that complement is connected, the poles can be pushed to infinity and the rational functions replaced by polynomials.

The obstruction is visible on the unit circle. There $\int_{|z|=1} z^{-1} dz = 2\pi i$, while the integral of every polynomial vanishes by Cauchy's theorem in the disc. A uniform limit preserves the integral, so $1/z$ is not a uniform limit of polynomials on the circle. The complement of the circle in the extended plane is disconnected: it has the disc and the exterior as separate pieces.

The approximation by rationals comes from a square-grid form of the integral formula. Let f be holomorphic on an open set containing K . Choose the mesh smaller than the distance from K to the complement, up to a fixed factor less than $1/\sqrt{2}$, and let $Q_1, \dots, Q_M$ be the closed squares of the grid that meet K , as in Figure 4.6. For z interior to one of those squares and off every boundary, the integral formula on that square returns $f(z)$, and the integral of $f(\zeta)/(\zeta - z)$ over every other square returns zero. Summing over the squares, each internal edge is traversed twice in opposite directions and cancels. The surviving edges $\gamma_1, \dots, \gamma_N$ lie in the complement of K , and

$$f(z) = \sum_{n=1}^N \frac{1}{2\pi i} \int_{\gamma_n} \frac{f(\zeta)}{\zeta - z} d\zeta. \quad (4.31)$$

Continuity extends the identity to all of K .

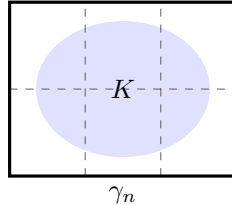


Figure 4.6: The squares that meet K . Internal edges cancel in pairs, and the surviving boundary, drawn heavily, is the chain γ_n in (4.31). Each of those edges lies off K , so the integrand $1/(\zeta - z)$ is holomorphic for $z \in K$.

Parametrize one segment γ by $\gamma(t)$ for $t \in [0, 1]$. The integrand $f(\gamma(t))\gamma'(t)/(\gamma(t) - z)$ is continuous in $(z, t) \in K \times [0, 1]$. Its Riemann sums are rational functions of z with poles on γ , and uniform continuity makes the sums converge uniformly on K . This is the rational approximation.

If the complement of K is connected, each pole can be moved to infinity. A rational function with a single pole at $z_0 \notin K$ is a polynomial in $1/(z - z_0)$. Choose a point z_1 outside a large disc that contains K . For $|z| < |z_1|$,

$$\frac{1}{z - z_1} = - \sum_{n=0}^{\infty} \frac{z^n}{z_1^{n+1}}, \quad (4.32)$$

uniformly on K , so $1/(z - z_1)$ and all of its powers are uniform limits of polynomials on K . Connect z_0 to z_1 by a curve γ in the complement, and set ρ equal to half the distance from K to γ . Pick points $w_0 = z_0, \dots, w_\ell = z_1$ along γ with $|w_{j+1} - w_j| < \rho$. For $z \in K$ one has $|z - w_j| \geq 2\rho$, hence

$$\frac{1}{z - w_{j+1}} = \sum_{k=0}^{\infty} \frac{(w_{j+1} - w_j)^k}{(z - w_j)^{k+1}}, \quad (4.33)$$

and the ratio of successive terms is strictly less than $1/2$. Thus $1/(z - w_{j+1})$ is a uniform limit of polynomials in $1/(z - w_j)$. Starting from z_1 and walking back to z_0 shows that $1/(z - z_0)$ is a uniform limit of polynomials on K .

4.8 Harmonic conjugates and the Poisson kernel

The integral formula at the center of the circle is a mean. Parametrize $\zeta = a + re^{i\theta}$ in (4.17):

$$f(a) = \frac{1}{2\pi} \int_0^{2\pi} f(a + re^{i\theta}) d\theta. \quad (4.34)$$

The real part is the spherical mean of u . Section 3.2 derived the same identity for a harmonic function in every dimension from the divergence theorem. In the plane it is the real part of (4.17). The same bound applied to $|f|$ is the maximum modulus principle. From (4.34),

$$|f(a)| \leq \frac{1}{2\pi} \int_0^{2\pi} |f(a + re^{i\theta})| d\theta \leq \max_{|\zeta - a| = r} |f|. \quad (4.35)$$

If $|f|$ attains an interior maximum at a , equality holds in the average, so $|f|$ equals $|f(a)|$ throughout the disc. Then $u^2 + v^2$ is constant, and its derivatives vanish: $uu_x + vv_x = 0$ and $uu_y + vv_y = 0$. Insert $v_x = -u_y$ and $v_y = u_x$. The system becomes

$$\begin{pmatrix} u & -v \\ v & u \end{pmatrix} \begin{pmatrix} u_x \\ u_y \end{pmatrix} = 0. \quad (4.36)$$

The determinant is $|f|^2$. Where $|f| > 0$ the only solution is $u_x = u_y = 0$, hence $f' = 0$. Where the maximum is 0, one has $f = 0$. Thus f is constant on the disk. If f is holomorphic on a bounded domain and continuous up to the boundary, $|f(z)|$ is at most the maximum of $|f|$ on the boundary: a larger interior value would be an interior maximum, and the disc argument would force f constant, in which case $|f|$ equals its boundary values. The maximum principle for a harmonic function is the same averaging applied directly to u , as already proved from the mean-value property in Section 3.2.

The two-dimensional Newtonian potential is the real part of a logarithm. On the plane cut along the nonpositive real axis, the principal branch $\text{Log } z = \log |z| + i \text{Arg } z$, with $\text{Arg} \in (-\pi, \pi)$, has derivative $1/z$. Differentiating $e^{\text{Log } z} = z$ gives the derivative. Therefore

$$\log |z| = \text{Re } \text{Log } z, \quad \frac{d}{dz} \text{Log } z = \frac{1}{z}. \quad (4.37)$$

For $n = 2$ one has $n\alpha_2 = 2\pi$, so the fundamental solution of Section 3.2 is $\Phi(z) = -(1/(2\pi)) \log |z| = -(1/(2\pi)) \operatorname{Re} \operatorname{Log} z$. The flux computation that produced $-\Delta\Phi = \delta$ is the winding of $1/z$. On $|z| = \varepsilon$,

$$\int_{|z|=\varepsilon} \frac{dz}{z} = 2\pi i, \quad (4.38)$$

while the real flux of $\nabla\Phi$ through that circle equals 1. The change in Arg around the origin is 2π , and dividing by 2π normalizes the flux.

A harmonic function on a disc is the real part of a holomorphic function. Let $\Delta u = 0$ on a convex open set, with u twice continuously differentiable, and fix a base point (x_0, y_0) . Define

$$v(x, y) = \int_{x_0}^x -u_y(s, y_0) ds + \int_{y_0}^y u_x(x, t) dt. \quad (4.39)$$

The fundamental theorem gives $v_y = u_x$. Differentiating under the integral and using $u_{xx} = -u_{yy}$ gives

$$v_x = -u_y(x, y_0) + \int_{y_0}^y u_{xx}(x, t) dt = -u_y(x, y). \quad (4.40)$$

Thus v satisfies (2.45), and both partials are continuous. The Cauchy–Riemann criterion of Section 4.7 says that $u + iv$ is holomorphic. Adding a constant to v produces every other conjugate.

The Poisson kernel of the disc is the real form of (4.17). Let f be holomorphic in a neighborhood of the closed disc $|\zeta| \leq r$, and let $0 < |z| < r$. The point $z^* = r^2/\bar{z}$ lies outside the disc, so the integral of $f(\zeta)/(\zeta - z^*)$ over the circle vanishes. Subtract it from (4.17). With $\zeta = re^{i\phi}$ one has $d\zeta = i\zeta d\phi$, and

$$\zeta \left(\frac{1}{\zeta - z} - \frac{1}{\zeta - z^*} \right) = \frac{r^2 - |z|^2}{|\zeta - z|^2}. \quad (4.41)$$

The difference of the summands is $(z - z^*)/((\zeta - z)(\zeta - z^*))$, and $z - z^* = (|z|^2 - r^2)/\bar{z}$, so the left side of (4.41) equals $\zeta(|z|^2 - r^2)/((\zeta - z)(\bar{z}\zeta - r^2))$. On $|\zeta| = r$ one has $r^2 = \zeta\bar{\zeta}$, hence $\bar{z}\zeta - r^2 = \zeta(\bar{z} - \bar{\zeta})$. One factor of ζ cancels. The remaining denominator is $(\zeta - z)(\bar{z} - \bar{\zeta}) = -|\zeta - z|^2$, which turns the numerator $|z|^2 - r^2$ into $r^2 - |z|^2$. At $z = 0$ the weight equals 1 and agrees with the circle mean. Therefore

$$f(z) = \frac{1}{2\pi} \int_0^{2\pi} f(re^{i\phi}) \frac{r^2 - |z|^2}{|re^{i\phi} - z|^2} d\phi. \quad (4.42)$$

The weight is real. Taking real parts gives the Poisson integral of the boundary values of $u = \operatorname{Re} f$. In the notation of (3.85) with $n = 2$, one has $n\alpha_2 = 2\pi$ and the surface measure is $r d\phi$, so the weight times $d\phi$ is $K dS$. The kernel obtained here by cancelling the exterior pole against the Cauchy kernel is the kernel obtained in Section 3.3.3 by cancelling the Newtonian potential at the inverse point.

A bounded harmonic function on the whole plane is constant. Translate (4.42) so that the disc of radius r is centered at the origin, and write the weight as N/D with $N = r^2 - |z|^2$ and $D = |\zeta - z|^2$. At the origin, $N = r^2$, $D = r^2$, $\partial_x N = 0$, and $\partial_x D = -2r \cos \phi$. The quotient rule gives $\partial_x(N/D) = 2 \cos \phi / r$ at the origin. If $|u| \leq M$ on the disc,

$$|\partial_x u(0)| \leq \frac{1}{2\pi} \int_0^{2\pi} M \frac{2|\cos \phi|}{r} d\phi \leq \frac{2M}{r}, \quad (4.43)$$

and the same bound holds for $\partial_y u$. For a function bounded on $\mathbb{R}^2$ every r is allowed. Letting $r \rightarrow \infty$ forces both partial derivatives to vanish.

4.9 Conformal maps

A map ϕ between plane domains is conformal when it is holomorphic and $\phi' \neq 0$. The Jacobian (2.47) says what this condition means geometrically. At a point where $\phi' = |\phi'|e^{i\alpha}$, the differential multiplies every tangent vector by $|\phi'|$ and rotates it by α . Two curves meeting at a point have tangents w_1 and w_2 . Their images have tangents $\phi'w_1$ and $\phi'w_2$. Adding α to both arguments leaves the difference of the arguments unchanged, so the oriented angle between the curves is preserved.

The local inverse is holomorphic. The Jacobian determinant equals $|\phi'|^2$. Where $\phi' \neq 0$ the real inverse-function theorem supplies a continuously differentiable local inverse. Differentiating $\phi^{-1}(\phi(z)) = z$ gives $(\phi^{-1})'(\phi(z))\phi'(z) = 1$, so the complex derivative of the inverse exists and equals $1/\phi'(z)$. The chain rule also gives $(\phi \circ \psi)' = \phi'(\psi)\psi'$. A composition of conformal maps is conformal where the intermediate derivative is nonzero.

Harmonic functions are preserved. Let $\phi = \xi + i\eta$ be holomorphic, let $u(\xi, \eta)$ be twice continuously differentiable, and set $U = u \circ \phi$. The chain rule expands to

$$\begin{aligned} U_{xx} + U_{yy} &= u_{\xi\xi}(\xi_x^2 + \xi_y^2) + u_{\eta\eta}(\eta_x^2 + \eta_y^2) \\ &\quad + 2u_{\xi\eta}(\xi_x\eta_x + \xi_y\eta_y) + u_{\xi}\Delta\xi + u_{\eta}\Delta\eta. \end{aligned} \quad (4.44)$$

Cauchy–Riemann gives $\xi_x = \eta_y$ and $\xi_y = -\eta_x$, so $\xi_x^2 + \xi_y^2 = \eta_x^2 + \eta_y^2 = |\phi'|^2$ and $\xi_x\eta_x + \xi_y\eta_y = 0$. Each component of a holomorphic function is harmonic, so $\Delta\xi = \Delta\eta = 0$. The identity that remains is

$$\Delta(u \circ \phi) = |\phi'|^2 \Delta u. \quad (4.45)$$

Wherever $\phi' \neq 0$, U is harmonic if and only if u is harmonic. Solving Laplace’s equation on a plane domain is the same problem as solving it on the image, with boundary values carried along ϕ .

The Möbius transformation that sends the upper half-plane to the unit disc is

$$\phi(z) = \frac{z-i}{z+i}, \quad \phi'(z) = \frac{2i}{(z+i)^2}. \quad (4.46)$$

The derivative is nonzero for $z \neq -i$. If z is real then $|z-i| = |z+i|$, so ϕ sends the real axis to the unit circle. If $\text{Im } z > 0$ then z is closer to i than to $-i$, so $|\phi(z)| < 1$, and $\phi(i) = 0$. The inverse is $\phi^{-1}(w) = i(1+w)/(1-w)$. Suppose U is harmonic on the disc, with boundary value g on the circle. Then $u = U \circ \phi$ is harmonic on the upper half-plane by (4.45), and the boundary value of u at a real point t is $g(\phi(t))$. The function U is the Poisson integral (4.42) of g . Composing with ϕ solves the half-plane problem whose boundary values are $g \circ \phi$. Composing a half-plane solution with ϕ^{-1} solves the Dirichlet problem on the disc. The two kernels of Section 3.3.3 are one representation transported by (4.46).

A power map opens a corner. On the sector $0 < \arg z < \alpha$, with $0 < \alpha \leq 2\pi$, choose the argument continuously with values in $(0, \alpha)$ and set

$$\phi(z) = |z|^{\pi/\alpha} \exp(i(\pi/\alpha) \arg z). \quad (4.47)$$

The map is holomorphic on the sector and has nonzero derivative. Arguments are multiplied by π/α , so the image is the upper half-plane, as in Figure 4.7. The function $\pi^{-1} \arg \phi(z)$ equals $\alpha^{-1} \arg z$. It is harmonic on the sector, equals 0 on the positive real axis, and equals 1 on the ray of argument α . A prescribed boundary function on the real line, solved by (3.81) and composed with ϕ , solves the Dirichlet problem on the sector.

Riemann’s mapping theorem is the existence statement that makes this a general method. If Ω is a simply connected plane domain, $\Omega \neq \mathbb{C}$, and $z_0 \in \Omega$, there is a conformal bijection of Ω

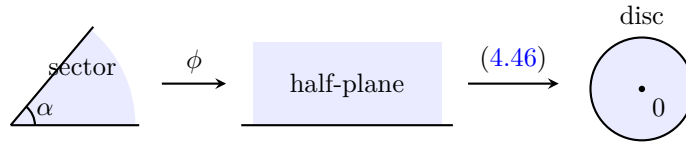


Figure 4.7: Two conformal changes of domain for Laplace's equation. The power (4.47) multiplies arguments by π/α and opens the sector onto the upper half-plane. The Möbius map (4.46) then sends that half-plane to the disc, with i going to 0. Boundary values travel with the map, and (4.45) keeps the solution harmonic.

onto the unit disc sending z_0 to 0, unique once the argument of its derivative at z_0 is fixed. The proof is a compactness argument for a family of holomorphic functions and is not given here. The Dirichlet problem uses (4.45) together with (4.42): boundary values on $\partial\Omega$ are read on the circle, the disc is solved by the Poisson integral, and the conformal map brings the solution back. The maps (4.46) and (4.47) are the cases in which the bijection is elementary. In dimension $d > 2$ a conformal map of a domain is the restriction of a Möbius transformation, by the spatial form of Liouville's theorem, and there is no analogue of Riemann's theorem. The Poisson kernel of a ball in higher dimension remains the one constructed by the inverse point in Section 3.3.3.

The scalar integral (4.17) evaluates a holomorphic function at a point. Chapter 11 writes the same integral for an operator, with $1/(z - a)$ replaced by the resolvent $(zI - A)^{-1}$. The operator in that formula is the matrix built in Chapter 5.

Chapter 5

Discretization and the linear-algebra reduction

Chapter 2 stated the continuous problem: an unknown function, a type, and data that make the problem well-posed. Chapter 3 writes the solution, on a simple domain, as a convolution against a kernel. A quantum algorithm never sees that function. It sees a matrix $A \in \mathbb{C}^{N \times N}$ (or a short family of such matrices) together with a vector that encodes forcing or initial data. The present chapter is the map from the kernel to that matrix [29, 33, 34].

Two reductions dominate everything that follows.

- An elliptic boundary-value problem discretizes to a linear system $A\mathbf{u} = \mathbf{f}$.
- A time-dependent problem, after discretization in space only, becomes a semidiscrete ODE $\mathbf{u}'(t) = A\mathbf{u}(t) + \mathbf{f}(t)$.

Boundary conditions are not an afterthought: they decide which nodes are unknowns, which entries of A are modified, and whether A is invertible. The classification of Section 2.5 is inherited by the matrices below: the stencil is the same second-order operator, written on a grid.

Section 5.1 records the mean-value theorems, Taylor expansion in one and two variables, and the norms in which errors are measured. Section 5.2 is the iteration those theorems justify. Interpolation, approximation, and quadrature then build the polynomials used on the grid, and the time step applies the same calculus to $\mathbf{u}' = A\mathbf{u}$.

5.1 Preliminaries

The constructions of this chapter rest on a short list of calculus facts and on the norms in which their errors are measured [32].

5.1.1 Mean-value theorems

Let f be continuous on a finite interval $[a, b]$, and write $m = \inf_{[a, b]} f$ and $M = \sup_{[a, b]} f$. The intermediate-value theorem says that every value $c \in [m, M]$ is attained: $f(x_0) = c$ for some $x_0 \in [a, b]$. In particular, $f(a)f(b) < 0$ forces a root in (a, b) . That is the existence statement of the bisection method in Section 5.2.

If f is also differentiable on (a, b) , the mean-value theorem supplies a point $\xi \in (a, b)$ with

$$f(b) - f(a) = f'(\xi)(b - a). \quad (5.1)$$

Applied to a differentiable function that vanishes at two points, (5.1) is Rolle's theorem: the derivative has a zero in between. Repeated application is the step from $n+2$ roots of (5.40) to the interpolation remainder (5.41).

If $w \geq 0$ is integrable on $[a, b]$ and f is continuous there, the integral mean-value theorem supplies a point $\xi \in [a, b]$ with

$$\int_a^b w(x) f(x) dx = f(\xi) \int_a^b w(x) dx. \quad (5.2)$$

The trapezoidal remainder in Section 5.5 is this identity for the weight $-(x-a)(x-a-h)$.

5.1.2 Taylor's formula

If u is infinitely differentiable at a point a , the Taylor series about a is

$$u(x) = \sum_{k=0}^{\infty} \frac{u^{(k)}(a)}{k!} (x-a)^k. \quad (5.3)$$

The series equals $u(x)$ only where it converges to u . A difference scheme uses the finite identity: if u is C^{n+1} between a and x , Taylor's theorem of order n gives some ξ strictly between a and x with

$$u(x) = \sum_{k=0}^n \frac{u^{(k)}(a)}{k!} (x-a)^k + \frac{u^{(n+1)}(\xi)}{(n+1)!} (x-a)^{n+1}. \quad (5.4)$$

For the exponential, $a = 0$ gives

$$e^x = \sum_{k=0}^n \frac{x^k}{k!} + \frac{e^\xi x^{n+1}}{(n+1)!}. \quad (5.5)$$

Section 5.4 returns to the degree-3 case and the size of the remainder on $[-1, 1]$. The interpolation error of Section 5.3, the trapezoidal and Simpson remainders of Section 5.5, and the truncation term of every stencil in Section 5.6 are (5.4), specialized to a polynomial, an integral, or a grid increment.

5.1.3 Taylor's formula in two variables

Let $f \in C^{n+1}(D)$ on a domain $D \subset \mathbb{R}^2$ containing the segment from (x_0, y_0) to (x, y) . Set $\xi = x - x_0$ and $\eta = y - y_0$. Then, for some $\theta \in [0, 1]$,

$$\begin{aligned} f(x, y) &= f(x_0, y_0) + \sum_{m=1}^n \frac{1}{m!} (\xi \partial_x + \eta \partial_y)^m f(x_0, y_0) \\ &\quad + \frac{1}{(n+1)!} (\xi \partial_x + \eta \partial_y)^{n+1} f(x_0 + \theta \xi, y_0 + \theta \eta). \end{aligned} \quad (5.6)$$

The case $n = 0$ is the mean-value theorem along that segment. The five-point Laplacian of Section 5.7.1 is the order-2 truncation of (5.6) at the four axial neighbours.

5.1.4 Norms

A norm satisfies $\|u\| \geq 0$, with equality only for $u = 0$; $\|\alpha u\| = |\alpha| \|u\|$; and $\|u + w\| \leq \|u\| + \|w\|$. On $\mathbb{R}^n$ and on continuous or square-integrable functions the examples used below are

$$\|x\|_\infty = \max_i |x_i|, \quad \|x\|_2 = \left(\sum_i x_i^2 \right)^{1/2}, \quad \|f\|_\infty = \max_{[a,b]} |f|, \quad \|f\|_{L^2(\Omega)} = \left(\int_\Omega |f|^2 dx \right)^{1/2}. \quad (5.7)$$

The weighted norm of Section 5.4 is the L^2 norm for the measure $w(x) dx$. A matrix norm induced by $\|\cdot\|_2$ is $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$. For a symmetric matrix this equals the spectral radius, so the norm of the Dirichlet matrix (5.76) is its largest eigenvalue and the condition number (5.78) is $\|A\|_2 \|A^{-1}\|_2$.

5.1.5 Sources of error

If x is an exact value and x_A an approximation, the absolute and relative errors are

$$x - x_A, \quad \frac{x - x_A}{x}. \quad (5.8)$$

The discrepancy $x - x_A$ may come from the modeling that replaced a physical system by a PDE, from a mistake in arithmetic or in a program, from uncertain data, from the floating-point representation $fl(x)$ (rounding or chopping, overflow or underflow), or from a mathematical truncation. The truncation is the object of the rest of this chapter: it is the local residual (5.123). Section 5.13.5 separates that contribution from the tolerance of the later solver and from the sampling error of the observable that the solver returns.

5.2 Rootfinding

A root of $f(x) = 0$, a nonlinear system, and an unconstrained minimum are computed by a sequence $x_0, x_1, x_2, \dots$ [32]. The sequence converges to α with order $p \geq 1$ when $|\alpha - x_{n+1}| \leq c |\alpha - x_n|^p$ for some $c > 0$ and all large n . For $p = 1$ one requires $c < 1$, and c is the linear rate. Each correction below is built from (5.1) or (5.4). On a grid the same correction is a linear system, which is why Chapter 15 meets these iterations again as QLSP instances.

5.2.1 Bisection

Suppose f is continuous on $[a, b]$ and $f(a)f(b) < 0$. The intermediate-value theorem places at least one root in (a, b) . Set $c = (a + b)/2$. If $f(b)f(c) < 0$, a root lies in $[c, b]$ and a is replaced by c ; otherwise b is replaced by c . The interval bisected at step n has length $2^{1-n}(b - a)$. The midpoint c_n of that interval lies at most half a length from α , so

$$|\alpha - c_n| \leq 2^{-n}(b - a). \quad (5.9)$$

Each step multiplies the error bound by $1/2$, which is linear convergence of rate $1/2$. The interval retained at that step still contains a root.

5.2.2 Newton's method

Assume $f \in C^2$ on an interval about a root α , with $f'(\alpha) \neq 0$. The expansion (5.4) of order 1 about the current guess x_n is

$$f(x) = f(x_n) + (x - x_n)f'(x_n) + \frac{(x - x_n)^2}{2!} f''(\xi), \quad (5.10)$$

where ξ lies between x and x_n . Substitute the root $x = \alpha$. Then $f(\alpha) = 0$, and

$$0 = f(x_n) + (\alpha - x_n)f'(x_n) + \frac{(\alpha - x_n)^2}{2!} f''(\xi). \quad (5.11)$$

Move $f(x_n)$ to the left and divide by $f'(x_n)$:

$$\alpha = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{(\alpha - x_n)^2}{2!} \frac{f''(\xi)}{f'(x_n)}. \quad (5.12)$$

Newton's method keeps the first correction and drops the quadratic term,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \quad (5.13)$$

Subtract (5.13) from (5.12). Since $2! = 2$,

$$\alpha - x_{n+1} = -(\alpha - x_n)^2 \frac{f''(\xi)}{2f'(x_n)}. \quad (5.14)$$

If $f' \neq 0$ throughout a neighborhood of α and x_0 is close enough that every later x_n remains there, the quotient $f''(\xi)/(2f'(x_n))$ stays bounded by some constant c . Then $|\alpha - x_{n+1}| \leq c|\alpha - x_n|^2$, so the convergence has order 2. A convergent step roughly doubles the number of correct digits. The correction uses the value of $f'(x_n)$. A distant x_0 can send the iterates away from α , and (5.13) is undefined at a point where f' vanishes. The next subsection builds the same correction from values of f alone.

5.2.3 The secant method

Replace $f'(x_n)$ in (5.13) by the slope of the chord through the two most recent values:

$$x_{n+1} = x_n - f(x_n) \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}. \quad (5.15)$$

The slope is the first divided difference $f[x_{n-1}, x_n] = (f(x_n) - f(x_{n-1}))/(x_n - x_{n-1})$, so (5.15) reads $x_{n+1} = x_n - f(x_n)/f[x_{n-1}, x_n]$.

The error formula is the Taylor expansion (5.10) with that chord in place of the tangent. Form the difference quotient that runs from x_n to the root,

$$f[x_n, \alpha] = \frac{f(\alpha) - f(x_n)}{\alpha - x_n}. \quad (5.16)$$

Clearing the denominator produces the exact linear rearrangement

$$f(\alpha) = f(x_n) + f[x_n, \alpha](\alpha - x_n). \quad (5.17)$$

The mean-value theorem (5.1) supplies a point between x_n and α at which $f[x_n, \alpha] = f'(\xi)$, so (5.17) is the linear part of (5.10) at that intermediate point. The slope $f[x_n, \alpha]$ still depends on the unknown value $f(\alpha)$. The difference between this slope and the known slope $f[x_{n-1}, x_n]$, divided by the gap of the free endpoints, is the second divided difference

$$f[x_{n-1}, x_n, \alpha] = \frac{f[x_n, \alpha] - f[x_{n-1}, x_n]}{\alpha - x_{n-1}}. \quad (5.18)$$

Solving for $f[x_n, \alpha]$ gives

$$f[x_n, \alpha] = f[x_{n-1}, x_n] + (\alpha - x_{n-1}) f[x_{n-1}, x_n, \alpha]. \quad (5.19)$$

Insert (5.19) into (5.17). The correction multiplies $(\alpha - x_n)$ by $(\alpha - x_{n-1})$, and

$$f(\alpha) = f(x_n) + f[x_{n-1}, x_n](\alpha - x_n) + f[x_{n-1}, x_n, \alpha](\alpha - x_n)(\alpha - x_{n-1}). \quad (5.20)$$

The new coefficient is the quadratic term of Taylor's formula. Let $p(t) = f(x_n) + f[x_{n-1}, x_n](t - x_n)$ and set

$$\phi(t) = f(t) - p(t) - f[x_{n-1}, x_n, \alpha](t - x_n)(t - x_{n-1}). \quad (5.21)$$

Then p matches f at x_n and at x_{n-1} , and (5.20) forces $\phi(\alpha) = 0$, so ϕ has three roots. Two applications of Rolle's theorem produce a point ξ with $\phi''(\xi) = 0$. The second derivative of p vanishes, and the second derivative of $(t - x_n)(t - x_{n-1})$ equals 2, hence $\phi''(\xi) = f''(\xi) - 2f[x_{n-1}, x_n, \alpha] = 0$. Therefore

$$f[x_{n-1}, x_n, \alpha] = \frac{f''(\xi)}{2!}, \quad (5.22)$$

and substitution into (5.20) recovers the expansion

$$f(\alpha) = f(x_n) + f[x_{n-1}, x_n](\alpha - x_n) + \frac{(\alpha - x_n)(\alpha - x_{n-1})}{2!} f''(\xi). \quad (5.23)$$

The constant term is $f(x_n)$, the linear coefficient is the chord slope $f[x_{n-1}, x_n]$, and the remainder carries the product $(\alpha - x_n)(\alpha - x_{n-1})$ in front of $f''(\xi)/2!$. This is (5.11) read on the chord through x_{n-1} and x_n .

Since α is a root, the left side of (5.20) is 0. The secant equation (5.15) is the same identity with the last term omitted and with α replaced by x_{n+1} , so $f(x_n) = f[x_{n-1}, x_n](x_n - x_{n+1})$. Substitute that expression for $f(x_n)$ into (5.20):

$$0 = f[x_{n-1}, x_n](\alpha - x_{n+1}) + f[x_{n-1}, x_n, \alpha](\alpha - x_n)(\alpha - x_{n-1}). \quad (5.24)$$

Divide by the first divided difference:

$$\alpha - x_{n+1} = -(\alpha - x_n)(\alpha - x_{n-1}) \frac{f[x_{n-1}, x_n, \alpha]}{f[x_{n-1}, x_n]}. \quad (5.25)$$

If f , f' , and f'' are continuous near α , if $f'(\alpha) \neq 0$, and if x_0 and x_1 are close to α , the product of the two previous errors produces convergence of order $(1 + \sqrt{5})/2$. Each step evaluates f once; the value $f(x_{n-1})$ is already known.

5.2.4 Fixed-point iteration

Any root problem may be rewritten as $x = g(x)$. The iteration is $x_{n+1} = g(x_n)$, and a solution of $x = g(x)$ is a fixed point of g . Suppose g is continuous on $[a, b]$, maps the interval into itself, and satisfies $|g(x) - g(y)| \leq \lambda|x - y|$ with $0 < \lambda < 1$. A fixed point exists because the inclusion forces $g(a) \geq a$ and $g(b) \leq b$, hence $(g(a) - a)(g(b) - b) \leq 0$, and the intermediate-value theorem supplies a zero of $g(x) - x$. Uniqueness follows from the contraction, since two fixed points would be at distance zero. The error itself contracts,

$$|\alpha - x_{n+1}| = |g(\alpha) - g(x_n)| \leq \lambda|\alpha - x_n|, \quad (5.26)$$

and therefore $|\alpha - x_n| \leq \lambda^n |\alpha - x_0|$. The same Lipschitz bound on successive steps gives $|x_{k+1} - x_k| \leq \lambda^k |x_1 - x_0|$. Telescoping $\alpha - x_0 = \sum_{k=0}^{\infty} (x_{k+1} - x_k)$ and summing the geometric series produces $|\alpha - x_0| \leq |x_1 - x_0|/(1 - \lambda)$, hence

$$|\alpha - x_n| \leq \frac{\lambda^n}{1 - \lambda} |x_1 - x_0|. \quad (5.27)$$

The iterates converge for every $x_0 \in [a, b]$. If in addition $g \in C^1([a, b])$ and $\lambda = \max_{[a, b]} |g'|$, the mean-value theorem supplies $\alpha - x_{n+1} = g'(\xi_n)(\alpha - x_n)$ with ξ_n between α and x_n . Dividing by $\alpha - x_n$ and sending $n \rightarrow \infty$ yields $(\alpha - x_{n+1})/(\alpha - x_n) \rightarrow g'(\alpha)$.

5.2.5 Aitken extrapolation

For a linearly convergent fixed-point iteration with $0 < |g'(\alpha)| < 1$, the ratios $\lambda_n = (x_n - x_{n-1})/(x_{n-1} - x_{n-2})$ tend to $g'(\alpha)$. Impose the two-term model $\alpha - x_n = \lambda(\alpha - x_{n-1})$ and $\alpha - x_{n-1} = \lambda(\alpha - x_{n-2})$ with the same λ . Subtraction gives

$$x_{n-1} - x_n = \lambda(x_{n-2} - x_{n-1}), \quad \lambda = \frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}}. \quad (5.28)$$

Rewrite $\alpha - x_{n-1} = (\alpha - x_n) + (x_n - x_{n-1})$ inside the first model equation:

$$\alpha - x_n = \lambda(\alpha - x_n) + \lambda(x_n - x_{n-1}), \quad \alpha = x_n + \frac{\lambda}{1 - \lambda}(x_n - x_{n-1}). \quad (5.29)$$

Insert $\lambda = \lambda_n$. Then

$$\frac{\lambda}{1 - \lambda}(x_n - x_{n-1}) = -\frac{(x_n - x_{n-1})^2}{(x_n - x_{n-1}) - (x_{n-1} - x_{n-2})}, \quad (5.30)$$

and Aitken's formula is the resulting approximation of α ,

$$\hat{x}_n = x_n - \frac{(x_n - x_{n-1})^2}{(x_n - x_{n-1}) - (x_{n-1} - x_{n-2})}. \quad (5.31)$$

The iteration to which the formula is applied is linearly convergent when

$$\lim_{n \rightarrow \infty} \frac{\alpha - x_{n+1}}{\alpha - x_n} = \mu, \quad 0 < |\mu| < 1. \quad (5.32)$$

For a fixed-point iteration the limit equals $g'(\alpha)$. Write $\rho_n = (\alpha - x_n)/(\alpha - x_{n-1})$, so that $\rho_n \rightarrow \mu$. The steps that enter (5.31) are then

$$x_n - x_{n-1} = (\alpha - x_{n-1})(1 - \rho_n), \quad x_{n-1} - x_{n-2} = (\alpha - x_{n-1}) \frac{1 - \rho_{n-1}}{\rho_{n-1}}. \quad (5.33)$$

Substitute (5.33) into (5.31). The common factor $\alpha - x_{n-1}$ cancels in the fraction, and the numerator that remains is the discrepancy of two successive ratios:

$$\alpha - \hat{x}_n = (\alpha - x_{n-1}) \frac{\rho_{n-1} - \rho_n}{2\rho_{n-1} - \rho_n\rho_{n-1} - 1}. \quad (5.34)$$

If every ratio equals the same constant μ , the numerator vanishes and $\hat{x}_n = \alpha$. In the limit $\rho_n \rightarrow \mu$ the numerator tends to 0 while the denominator tends to $2\mu - \mu^2 - 1 = -(1 - \mu)^2 \neq 0$, and since $(\alpha - x_n)/(\alpha - x_{n-1}) \rightarrow \mu \neq 0$,

$$\lim_{n \rightarrow \infty} \frac{\alpha - \hat{x}_n}{\alpha - x_n} = 0. \quad (5.35)$$

The extrapolated error is therefore asymptotically smaller than the error of the sequence from which it was built.

The denominator of (5.31) is a difference of successive steps. Once the ratios have settled near μ , that difference is about $(1 - \mu)/\mu$ times the latest step, so it tends to zero with the step. In floating-point arithmetic the two steps agree in their leading digits and the subtraction cancels them. Dividing by the contaminated denominator amplifies the rounding error, and the computed correction can then exceed the true distance from x_n to α . The extrapolation should be skipped when the denominator is no longer large compared with the size of the steps.

5.2.6 Systems

For $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$ the fixed-point form is $x = g(x)$. In two variables the mean-value theorem on each component assembles a 2×2 matrix G_n of partial derivatives of g , evaluated between α and x_n , with $\alpha - x_{n+1} = G_n(\alpha - x_n)$. Let $D \subset \mathbb{R}^m$ be closed, bounded, and convex, let $g(D) \subset D$, and let $G = Dg$. If $\lambda = \max_{x \in D} \|G(x)\|_\infty < 1$, where $\|\cdot\|_\infty$ is the maximum absolute row sum, then $x = g(x)$ has a unique solution in D and the iteration converges from every $x_0 \in D$. The scalar restriction $|g'| < 1$ is the case $m = 1$.

For two unknowns the same expansion as (5.10), applied to each component and stopped after the linear term, reads

$$\begin{aligned} 0 &\approx f_1(x_n, y_n) + (\alpha_1 - x_n) \partial_x f_1(x_n, y_n) + (\alpha_2 - y_n) \partial_y f_1(x_n, y_n), \\ 0 &\approx f_2(x_n, y_n) + (\alpha_1 - x_n) \partial_x f_2(x_n, y_n) + (\alpha_2 - y_n) \partial_y f_2(x_n, y_n). \end{aligned} \quad (5.36)$$

In matrix form this is $0 \approx f(x_n) + F(x_n)(\alpha - x_n)$, with $F = Df$ the Jacobian. Solving for the unknown increment $\delta = \alpha - x_n$ gives the practical step

$$F(x_n) \delta_{n+1} = -f(x_n), \quad x_{n+1} = x_n + \delta_{n+1}. \quad (5.37)$$

Locally $\|\alpha - x_{n+1}\|_\infty \leq B \|\alpha - x_n\|_\infty^2$ when x_0 is close to α . Freezing $F(x_n)$ for several steps reuses one factorization. A damped step $x_n + s\delta_{n+1}$ with $s \in (0, 1]$ chosen so that $\|f(x_n + s\delta_{n+1})\|_2$ decreases enlarges the set of x_0 from which the iteration converges. For a nonlinear PDE the unknown x_n is a grid vector and $F(x_n)$ is a sparse matrix of the kind built later in this chapter, so (5.37) is a QLSP instance.

5.2.7 Unconstrained optimization

A local minimizer α of a smooth function $f(x_1, \dots, x_m)$ satisfies $\nabla f(\alpha) = 0$. Newton's method on that system is

$$x_{n+1} = x_n - H(x_n)^{-1} \nabla f(x_n), \quad (5.38)$$

where $H_{ij} = \partial^2 f / \partial x_i \partial x_j$ is the Hessian. A descent method chooses a direction d_n along which f decreases and then minimizes $\phi(s) = f(x_n + sd_n)$ for $s \geq 0$. The steepest-descent choice is $d_n = -\nabla f(x_n)$. A quasi-Newton update builds an approximation of H from successive gradients. A constraint such as $x > 0$ can be moved into the objective by a barrier, for instance $F(x) + \delta |\log x|$, after which the problem is unconstrained and (5.38) applies.

5.3 Interpolation

Given $n+1$ distinct nodes $x_0, \dots, x_n$ and values $y_0, \dots, y_n$, a polynomial $p(x) = \sum_{j=0}^n a_j x^j$ matches the data when the coefficients solve the Vandermonde system $\sum_{j=0}^n a_j x_i^j = y_i$ [32]. There is exactly one such polynomial of degree at most n . If r_1 and r_2 both interpolate, then $r_1 - r_2$ has degree at most n and $n+1$ roots, so the difference vanishes. The Lagrange formula exhibits the polynomial:

$$p_n(x) = \sum_{i=0}^n y_i \ell_i(x), \quad \ell_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j}. \quad (5.39)$$

The factor ℓ_i equals 1 at x_i and equals 0 at the other nodes, so $p_n(x_i) = y_i$.

The pointwise error is (5.4) written as a product over the nodes. Let $\omega(t) = (t - x_0) \cdots (t - x_n)$, and fix an evaluation point x distinct from the nodes, with $f \in C^{n+1}$ on an interval containing the nodes and x . The auxiliary function

$$\phi(t) = f(t) - p_n(t) - \frac{f(x) - p_n(x)}{\omega(x)} \omega(t) \quad (5.40)$$

vanishes at $x_0, \dots, x_n$ and at x , hence has $n + 2$ roots. Rolle's theorem, applied $n + 1$ times, yields a point ξ with $\phi^{(n+1)}(\xi) = 0$. Differentiation of order $n + 1$ annihilates p_n and sends ω to $(n + 1)!$, so

$$f(x) - p_n(x) = \frac{\omega(x)}{(n + 1)!} f^{(n+1)}(\xi). \quad (5.41)$$

On a uniform partition of width h the same formula with $n = 1$ is applied on each panel separately. The resulting piecewise linear interpolant is $\sum_j y_j \varphi_j$, where the hat function

$$\varphi_j(x) = \begin{cases} (x - x_{j-1})/h, & x \in [x_{j-1}, x_j], \\ (x_{j+1} - x)/h, & x \in [x_j, x_{j+1}], \\ 0, & \text{otherwise,} \end{cases} \quad (5.42)$$

equals 1 at x_j and equals 0 at every other node. On one panel $|(x - x_j)(x - x_{j+1})| \leq h^2/4$, so (5.41) gives $|f - p| \leq (h^2/8) \max |f''|$. The interior hats vanish at the endpoints of the interval. They are the piecewise-linear basis of the space H_0^1 used in Section 5.11. A single polynomial of rising degree on equally spaced nodes can diverge: the interpolant of $1/(1 + x^2)$ on $[-5, 5]$ grows near the endpoints as the degree increases. The Chebyshev nodes of Section 5.4 cluster at those endpoints.

5.4 Polynomial approximation

Section 3.5 produced p_n , T_n , L_n , and H_n as series solutions. The numerical use of those functions is approximation in a weighted inner product [32]. Taylor's formula (5.4) is one polynomial approximation, and its error need not be spread evenly: the degree-3 expansion of e^x about 0 leaves the remainder $e^\xi x^4/24$, which on $[-1, 1]$ is largest at the endpoints. Weierstrass' theorem says a better polynomial always exists. If f is continuous on a finite interval $[a, b]$ and $\varepsilon > 0$, some polynomial p satisfies $|f(x) - p(x)| \leq \varepsilon$ throughout $[a, b]$.

The monomials are a poor basis in which to compute that polynomial. Least squares for $\sum_{j=0}^n c_j x^j$ against a weight w produces the Gram matrix

$$A_{ij} = \int_a^b w(x) x^{i+j} dx. \quad (5.43)$$

On $[0, 1]$ with $w \equiv 1$ this is a Hilbert matrix. Its columns are nearly dependent, and the coefficients move sharply under a small change in f . The orthogonal polynomials of Section 3.5 are the basis that removes the dependence.

Let $w > 0$ on (a, b) and set

$$(f, g) = \int_a^b w(x) f(x) g(x) dx, \quad \|f\|_2 = \sqrt{(f, f)}. \quad (5.44)$$

Cauchy-Schwarz says $|(f, g)| \leq \|f\|_2 \|g\|_2$. Gram-Schmidt applied to the monomials produces $\deg \phi_n = n$ and $(\phi_n, \phi_m) = 0$ for $n \neq m$, and one may normalize $\|\phi_n\|_2 = 1$. The constant term

and the inductive step are

$$\phi_0 = \left(\int_a^b w \right)^{-1/2}, \quad \psi_n = x^n + \sum_{j=0}^{n-1} c_{n,j} \phi_j, \quad c_{n,j} = -(x^n, \phi_j), \quad (5.45)$$

with $\phi_n = \psi_n / \|\psi_n\|_2$. On $[-1, 1]$ with $w \equiv 1$ this reproduces the normalized Legendre polynomials

$$\phi_0 = \frac{1}{\sqrt{2}}, \quad \phi_1 = \sqrt{\frac{3}{2}} x, \quad \phi_2 = \frac{1}{2} \sqrt{\frac{5}{2}} (3x^2 - 1), \quad (5.46)$$

and $(p_n, p_n) = 2/(2n+1)$. The Chebyshev weight is $1/\sqrt{1-x^2}$. The substitution $x = \cos \theta$ converts the inner product into

$$(T_n, T_m) = \int_0^\pi \cos(n\theta) \cos(m\theta) d\theta = \begin{cases} 0, & n \neq m, \\ \pi, & n = m = 0, \\ \pi/2, & n = m \geq 1. \end{cases} \quad (5.47)$$

The Laguerre weight is e^{-x} on $[0, \infty)$, and $\|L_n\|_2 = 1$. The oscillator weight is e^{-x^2} on $\mathbb{R}$, and $\int_{-\infty}^\infty e^{-x^2} H_n(x) H_m(x) dx = \sqrt{\pi} 2^n n! \delta_{nm}$.

The product $x\phi_n$ has degree $n+1$, and $(x\phi_n, \phi_j) = (\phi_n, x\phi_j) = 0$ for every $j \leq n-2$. Hence $x\phi_n$ lies in $\text{span}\{\phi_{n-1}, \phi_n, \phi_{n+1}\}$,

$$x\phi_n = a_n \phi_{n+1} + b_n \phi_n + c_n \phi_{n-1}. \quad (5.48)$$

The recurrences (3.138), (3.141), (3.147), and (3.152) are this identity for the four normalizations. A polynomial f of degree m is the finite sum

$$f = \sum_{n=0}^m \frac{(f, \phi_n)}{(\phi_n, \phi_n)} \phi_n, \quad (5.49)$$

so $(f, \phi_m) = 0$ whenever $\deg f \leq m-1$. Each ϕ_n has n distinct roots in (a, b) : fewer sign changes would produce a polynomial q of degree less than n with $(q, \phi_n) \neq 0$, contradicting the previous sentence. Those roots are the nodes of the Gaussian rule in Section 5.5. For the Chebyshev weight the nodes are $x_j = \cos((2j-1)\pi/(2n))$.

For an orthonormal family and $r = \sum_{j=0}^n a_j \phi_j$,

$$\|f - r\|_2^2 = \|f\|_2^2 - \sum_{j=0}^n (f, \phi_j)^2 + \sum_{j=0}^n ((f, \phi_j) - a_j)^2. \quad (5.50)$$

The minimum is unique at $a_j = (f, \phi_j)$,

$$r_n = \sum_{j=0}^n (f, \phi_j) \phi_j, \quad \|f - r_n\|_2^2 = \|f\|_2^2 - \|r_n\|_2^2. \quad (5.51)$$

If the family is only orthogonal, the coefficient is $(f, \phi_j)/(\phi_j, \phi_j)$, the same quotient as (3.128). Bessel's inequality and, on a finite interval, Parseval's identity are

$$\sum_{j=0}^n (f, \phi_j)^2 \leq \|f\|_2^2, \quad \|f\|_2^2 = \sum_{j=0}^\infty (f, \phi_j)^2. \quad (5.52)$$

The second equality uses Weierstrass' theorem to place a polynomial arbitrarily close to f , together with (5.49). The Laguerre and Hermite weights are supported on unbounded intervals, so the completeness identity above is the finite-interval statement; the coefficient formula and Bessel's inequality still hold for L_n and H_n .

5.5 Quadrature

Integrating an interpolant replaces a definite integral by a finite sum [32]. On a panel $[a, a+h]$ the linear interpolant integrates to the trapezoidal rule. The error is the integral of the $n = 1$ case of (5.41). The weight $(x-a)(x-a-h)$ does not change sign on the panel, and $\int_a^{a+h} (x-a)(x-a-h) dx = -h^3/6$. The integral mean-value theorem (5.2), applied to the nonnegative weight $-(x-a)(x-a-h)$, pulls one value of f'' out of the integral and leaves

$$\int_a^{a+h} f(x) dx = \frac{h}{2}(f(a) + f(a+h)) - \frac{h^3}{12}f''(\eta) \quad (5.53)$$

for some η between a and $a+h$. On the uniform nodes $x_j = a + jh$ with $h = (b-a)/n$, adding the panels produces the composite rule

$$\int_a^b f(x) dx = h \left(\frac{1}{2}f(x_0) + \sum_{j=1}^{n-1} f(x_j) + \frac{1}{2}f(x_n) \right) - \frac{(b-a)h^2}{12}f''(\eta). \quad (5.54)$$

Simpson's rule integrates the quadratic interpolant at the ends and the midpoint of a panel. Write the panel as $[c-h, c+h]$ and expand $f(c \pm h)$ by (5.4) through order 5. Odd powers cancel, and

$$f(c-h) + 4f(c) + f(c+h) = 6f(c) + h^2f''(c) + \frac{h^4}{12}f^{(4)}(c) + O(h^6). \quad (5.55)$$

Multiplication by $h/3$ produces $2hf(c) + (h^3/3)f''(c) + (h^5/36)f^{(4)}(c) + O(h^7)$. The integral of the same expansion is $2hf(c) + (h^3/3)f''(c) + (h^5/60)f^{(4)}(c) + O(h^7)$. The coefficients of f and of f'' agree, and $1/60 - 1/36 = -1/90$, so the rules differ by $-(h^5/90)f^{(4)}$ at leading order. The remainder form of (5.4) replaces that leading term by a fourth derivative at an intermediate point:

$$\int_{c-h}^{c+h} f(x) dx = \frac{h}{3}(f(c-h) + 4f(c) + f(c+h)) - \frac{h^5}{90}f^{(4)}(\eta). \quad (5.56)$$

The rule is exact for cubics. Summing panels of half-width h across $[a, b]$ multiplies the local factor $h^5/90$ by the number $(b-a)/(2h)$ of panels, and the composite error is $-(b-a)h^4/180$ times a fourth derivative.

The nodes that raise the degree of precision from $n-1$ to $2n-1$ are the roots of the orthogonal polynomial ϕ_n from Section 5.4. Divide $f = q + \phi_n r$ with $\deg q \leq n-1$. Whenever $\deg f \leq 2n-1$ one has $\deg r \leq n-1$, so $(r, \phi_n) = 0$ by (5.49). An interpolatory rule at those roots integrates q exactly and assigns integral zero to $w\phi_n r$. Therefore

$$\int_a^b w(x) f(x) dx = \sum_{j=1}^n w_j f(x_j) \quad (5.57)$$

for every polynomial f of degree at most $2n-1$, when x_j are the roots of ϕ_n . For the Legendre weight $w \equiv 1$ on $[-1, 1]$ the weights are $w_j = 2/((1-x_j^2)(p'_n(x_j))^2)$. For the Chebyshev weight they are the constant π/n , at the nodes $x_j = \cos((2j-1)\pi/(2n))$. The stiffness matrix, mass matrix, and load vector of Section 5.11 are integrals of products of the hats (5.42), and the Duhamel formula of Section 5.12 is an integral in time. A fully discrete scheme replaces each of those integrals by a sum of this form.

5.6 Grid functions and stencils

A finite-difference scheme never differentiates a formula. It replaces each partial derivative at a grid point by a fixed linear combination of neighbouring values—a *stencil*—and writes one algebraic equation per unknown. The weights are the solution of a Vandermonde system, as in Section 5.3, chosen so that (5.4) reproduces the target derivative. The classical references are LeVeque and the notes [33, 29]; the quantum cost model later reads off only the resulting sparsity s , the scaling $\alpha \sim \|A\|$, and the spectrum.

5.6.1 Expansion on the grid

Fix a uniform mesh of spacing $h > 0$, with nodes $x_j = jh$. A grid function U_j approximates $u(x_j)$. The weights below are chosen by truncating (5.4) at the nodes $x_j + \ell h$ [33, 29].

On the grid one takes $a = x_j$ and an increment $\delta \in h\mathbb{Z}$, so $x = x_j + \delta$. Then (5.4) is

$$u(x_j + \delta) = \sum_{k=0}^n \frac{u^{(k)}(x_j)}{k!} \delta^k + O(\delta^{n+1}). \quad (5.58)$$

The cases $\delta = \pm h$ are

$$u(x_j \pm h) = \sum_{k=0}^n \frac{u^{(k)}(x_j)}{k!} (\pm h)^k + O(h^{n+1}). \quad (5.59)$$

(The minus sign on odd k is just $(\pm h)^k$.) From here on, derivatives without an argument are evaluated at x_j . Adding or subtracting the two choices of sign in (5.59) isolates even or odd powers:

$$\begin{aligned} u(x_j + h) + u(x_j - h) &= 2 \sum_{m=0}^{\lfloor n/2 \rfloor} \frac{u^{(2m)}}{(2m)!} h^{2m} + O(h^{n+1}), \\ u(x_j + h) - u(x_j - h) &= 2 \sum_{m=0}^{\lfloor (n-1)/2 \rfloor} \frac{u^{(2m+1)}}{(2m+1)!} h^{2m+1} + O(h^{n+1}). \end{aligned} \quad (5.60)$$

Every centered stencil below is a linear combination of these two identities. One-sided stencils use (5.58) at several increments of one sign. Either way, the design problem is the same: choose weights so that unwanted powers of h cancel and the coefficient of the target derivative is 1.

5.6.2 Stencils as coefficient matching

A one-dimensional stencil with offsets $\mathcal{S} \subset \mathbb{Z}$ is an operator

$$(\delta_h u)(x_j) = \sum_{\ell \in \mathcal{S}} c_\ell u(x_j + \ell h). \quad (5.61)$$

Substitute (5.58) with $\delta = \ell h$ and collect powers of h . The coefficient of $u^{(k)}(x_j)$ is $(h^k/k!) \sum_{\ell} c_\ell \ell^k$. To have $\delta_h u = u^{(m)} + O(h^p)$ one therefore solves

$$\sum_{\ell \in \mathcal{S}} c_\ell \ell^k = \begin{cases} m!/h^m, & k = m, \\ 0, & k \neq m, \ 0 \leq k \leq m + p - 1. \end{cases} \quad (5.62)$$

This is the Vandermonde system of Section 5.3, with the stencil offsets in the role of the nodes. Distinct offsets make the coefficient vector unique once $|\mathcal{S}|$ and any symmetry imposed on the c_ℓ

have fixed how many moments are matched. The first unmatched moment $k = m + p$ produces the leading truncation error.

The standard first differences [33, 29] are the two-point solutions of (5.62) for $m = 1$:

$$(D_+U)_j = \frac{U_{j+1} - U_j}{h}, \quad (D_-U)_j = \frac{U_j - U_{j-1}}{h}, \quad (D_0U)_j = \frac{U_{j+1} - U_{j-1}}{2h}. \quad (5.63)$$

Forward and backward differences use $\mathcal{S} = \{0, 1\}$ and $\{-1, 0\}$; they match $k = 0, 1$ only, so $p = 1$ and $D_\pm u = u' \pm (h/2)u'' + O(h^2)$. The centered operator uses the odd identity in (5.60) with $\mathcal{S} = \{-1, 1\}$, which kills every even k automatically; the first surviving error is $k = 3$, hence

$$D_0u = u' + \frac{h^2}{6}u''' + O(h^4) = u' + O(h^2). \quad (5.64)$$

Symmetry, not the number of points, is what raises the order.

5.6.3 The three-point second-derivative stencil

For $m = 2$ the even identity in (5.60), truncated after $k = 2$ and $k = 4$, rearranges to

$$(D_+D_-u)_j = \frac{u(x_j - h) - 2u(x_j) + u(x_j + h)}{h^2} = u'' + \frac{h^2}{12}u^{(4)} + O(h^4). \quad (5.65)$$

Equivalently, $\mathcal{S} = \{-1, 0, 1\}$ with $c_{\pm 1} = h^{-2}$, $c_0 = -2h^{-2}$ solves (5.62) for $m = 2$ up to $p = 2$. In stencil notation

$$D_+D_- = \frac{1}{h^2} \begin{bmatrix} 1 & -2 & 1 \end{bmatrix}. \quad (5.66)$$

Poisson uses the positive operator $-u''$, i.e. $h^{-2}[-1 \ 2 \ -1]$. The $O(h^2)$ local truncation error is a statement about one grid equation, not yet about the global error of $A\mathbf{u} = \mathbf{f}$.

5.6.4 One-sided closures

Near a boundary a centered molecule in $\mathcal{S}$ exits the domain. The same system (5.62) on a one-sided set, e.g. $\mathcal{S} = \{0, -1, -2\}$ for $m = 1$ and $p = 2$, yields

$$u'(x) = \frac{1}{2h}(3u(x) - 4u(x - h) + u(x - 2h)) + O(h^2). \quad (5.67)$$

Order p on one side needs $p + 1$ nodes and typically destroys the symmetry that made A Hermitian. Equation (5.67) is the Neumann closure that does not introduce a ghost node.

5.6.5 Raising the order

If a consistent operator admits an even error expansion $L_h u = Lu + a h^{2q} u^{(2q+m)} + O(h^{2q+2})$, the same operator on a mesh of width $2h$ has leading coefficient $2^{2q}a$. The combination

$$\frac{2^{2q}L_h - L_{2h}}{2^{2q} - 1} = L + O(h^{2q+2}) \quad (5.68)$$

cancels that term (Richardson extrapolation). For the second difference (5.65) one has $q = 1$ and $a = u^{(4)}/12$, so $(4D_+D_- - (D_+D_-)_{2h})/3$ is fourth-order. Written on the grid it is the five-point stencil

$$u''(x_j) = \frac{-u_{j-2} + 16u_{j-1} - 30u_j + 16u_{j+1} - u_{j+2}}{12h^2} + O(h^4). \quad (5.69)$$

The matrix becomes pentadiagonal ($s = 5$ rather than $s = 3$); $\|A\|$ is still $\Theta(h^{-2})$. The algebraic type does not change. The same extrapolation in two dimensions, applied to axial versus diagonal molecules, is how the nine-point compact scheme of Section 5.7.2 is obtained.

5.6.6 From a stencil to a matrix row

Once a stencil is chosen, the global matrix is defined by placing that molecule at every unknown node. Index the unknowns by $p = 1, \dots, N$ after boundary conditions have been imposed. A stencil with weights $a_{p,q}$ at neighbours $q \in \mathcal{N}(p)$ produces the p -th equation

$$\sum_{q \in \mathcal{N}(p)} a_{p,q} U_q = f_p. \quad (5.70)$$

The left-hand side is exactly row p of A . Interior nodes that see only other unknowns contribute a purely local row. A neighbour that lies on a Dirichlet boundary is not an unknown: its value is moved to the right-hand side, which modifies $\mathbf{f}$ and shortens $\mathcal{N}(p)$. A Neumann or periodic neighbour remains an unknown and stays in A , possibly wrapping around the grid. This is the only mechanism by which boundary data enter the linear algebra; Sections 5.7–5.9 write it for each standard condition.

Two quantities are then fixed for every later quantum algorithm. The *sparsity* $s = \max_p |\mathcal{N}(p)|$ is the stencil width (three-point: $s = 3$; five-point Laplacian: $s = 5$; nine-point: $s = 9$). The *dimension* N is the number of retained unknowns, stored in $\lceil \log_2 N \rceil$ qubits. Both are properties of the stencil and the mesh, not of HHL or QSVT.

5.7 Dirichlet Poisson: how the matrix is assembled

The model elliptic problem on the unit interval is

$$-u''(x) = f(x), \quad 0 < x < 1, \quad u(0) = \alpha, \quad u(1) = \beta. \quad (5.71)$$

Take n interior nodes with $h = 1/(n+1)$ and $x_j = jh$ for $j = 0, \dots, n+1$. The boundary values $U_0 = \alpha$ and $U_{n+1} = \beta$ are *data*, not unknowns. The unknown vector is therefore the interior grid function

$$\mathbf{u} = (U_1, \dots, U_n)^\top \in \mathbb{R}^n. \quad (5.72)$$

At each interior node the centered scheme (5.65) reads

$$\frac{-U_{j-1} + 2U_j - U_{j+1}}{h^2} = f(x_j), \quad j = 1, \dots, n. \quad (5.73)$$

For $2 \leq j \leq n-1$ every neighbour is an unknown, and (5.73) is simply a row of a tridiagonal matrix. At $j = 1$ the neighbour U_0 is known, so it must be moved to the right-hand side:

$$\frac{2U_1 - U_2}{h^2} = f(x_1) + \frac{\alpha}{h^2}. \quad (5.74)$$

At $j = n$ the same happens with β :

$$\frac{-U_{n-1} + 2U_n}{h^2} = f(x_n) + \frac{\beta}{h^2}. \quad (5.75)$$

Collecting the interior equations gives $A\mathbf{u} = \mathbf{f}$ with

$$A = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} f(x_1) + \alpha/h^2 \\ f(x_2) \\ \vdots \\ f(x_{n-1}) \\ f(x_n) + \beta/h^2 \end{pmatrix}. \quad (5.76)$$

The matrix A does *not* contain the Dirichlet data. Those data appear only in $\mathbf{f}$, scaled by h^{-2} , which is the same factor that sits in front of A . Homogeneous Dirichlet conditions $\alpha = \beta = 0$ therefore leave A unchanged and simply sample f at the interior nodes.

If one had used the fourth-order stencil (5.69) instead, A would be pentadiagonal and the first two rows would come from a one-sided closure of (5.62); the problem is still $A\mathbf{u} = \mathbf{f}$.

The matrix in (5.76) is symmetric, tridiagonal, and positive definite (SPD). Positive-definiteness is the discrete maximum principle: if $A\mathbf{u} = 0$ then $\mathbf{u} = 0$, so the homogeneous Dirichlet problem has only the zero solution [33, 34]. Equivalently, $A = G^\top G$ for the forward difference G that already incorporates $U_0 = U_{n+1} = 0$. The same factorization is what Li uses to reduce the spectral scale from h^{-2} to h^{-1} by a first-order dilation [34].

The matrix (5.76) is Toeplitz—constant along each diagonal—but not circulant: the corners that would wrap U_1 to U_n are absent, because those neighbours were Dirichlet data. It is diagonalized by the discrete sine transform (DST). The eigenvectors and eigenvalues are [34]

$$(v^{(k)})_j = \sin\left(\frac{k\pi j}{n+1}\right), \quad \lambda_k(A) = \frac{4}{h^2} \sin^2\left(\frac{k\pi}{2(n+1)}\right), \quad k = 1, \dots, n, \quad (5.77)$$

with $h = 1/(n+1)$. Equivalently $\lambda_k = (2 - 2\cos(k\pi/(n+1)))/h^2$. The sine modes vanish at the fictitious endpoints $j = 0$ and $j = n+1$, which is why they are the Dirichlet eigenfunctions. Section 5.10 returns to this as the odd part of a Fourier series on a doubled grid.

Hence $\lambda_{\min}(A) = \Theta(1)$ on the unit interval and $\lambda_{\max}(A) = \Theta(h^{-2})$, so the condition number is

$$\kappa(A) = \frac{\lambda_{\max}}{\lambda_{\min}} = \Theta(h^{-2}) = \Theta(n^2). \quad (5.78)$$

This κ is the parameter that later dominates every quantum linear-system bound. The passage from the local truncation error $O(h^2)$ to a global $\max_j |U_j - u(x_j)| = O(h^2)$ is the stability argument of Section 5.13; a quantum solver does not remove that discretization error, it only changes how $A\mathbf{u} = \mathbf{f}$ is solved.

5.7.1 Two dimensions: the five-point stencil

On a Cartesian product grid the Laplacian splits as $\Delta u = u_{xx} + u_{yy}$. Applying the three-point molecule (5.66) once in x and once in y , and adding, produces the five-point stencil

$$(\Delta_h U)_{i,j} = \frac{U_{i-1,j} + U_{i+1,j} + U_{i,j-1} + U_{i,j+1} - 4U_{i,j}}{h^2} = \frac{1}{h^2} \begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix} U_{i,j}. \quad (5.79)$$

(The array is understood to sit on the 3×3 neighbourhood of (i, j) , with empty corners equal to zero.) It is the sum of the one-dimensional operator (5.65) in x and in y , hence the truncation of (5.6) along the axial steps, so the truncation error is the sum of the one-dimensional errors:

$$\Delta_h u = \Delta u + \frac{h^2}{12} (u_{xxxx} + u_{yyyy}) + O(h^4). \quad (5.80)$$

The cross derivative u_{xxyy} is missing from the $O(h^2)$ error, so the five-point operator is only second-order accurate and is not rotationally invariant at that order. The nine-point stencil of Section 5.7.2 restores the missing term.

For the Dirichlet problem $-\Delta u = f$ on the unit square, let $U_{i,j}$ approximate $u(ih, jh)$ at interior nodes $i, j = 1, \dots, n$, with $h = 1/(n+1)$, and stack the unknowns lexicographically,

$$\mathbf{u} = (U_{1,1}, \dots, U_{n,1}, U_{1,2}, \dots, U_{n,n})^\top \in \mathbb{R}^{n^2}. \quad (5.81)$$

The index of node (i, j) is $p = i + (j-1)n$. The five-point molecule at an interior node then couples p to $p \pm 1$ (the x -neighbours) and to $p \pm n$ (the y -neighbours). In matrix language this is the Kronecker sum of the one-dimensional factor $T = h^{-2} \text{tridiag}(-1, 2, -1)$,

$$A = I_n \otimes T + T \otimes I_n. \quad (5.82)$$

Each row has at most five nonzeros ($s = 5$). Explicitly, an interior row corresponding to (i, j) with $1 < i < n$ and $1 < j < n$ reads

$$\frac{-U_{i-1,j} - U_{i+1,j} - U_{i,j-1} - U_{i,j+1} + 4U_{i,j}}{h^2} = f(x_i, y_j). \quad (5.83)$$

If $i = 1$, the neighbour $U_{0,j}$ is a Dirichlet value g on the left side of the square and is moved to the right-hand side as g/h^2 , exactly as α/h^2 was in one dimension. The four sides of the square are four copies of that mechanism; A itself is independent of the boundary function g . Eigenvectors of (5.82) are tensor products of one-dimensional sine modes, with

$$\lambda_{k,\ell}(A) = \frac{4}{h^2} \left[\sin^2\left(\frac{k\pi}{2(n+1)}\right) + \sin^2\left(\frac{\ell\pi}{2(n+1)}\right) \right]. \quad (5.84)$$

The condition number remains $\kappa(A) = \Theta(h^{-2})$ in any fixed dimension d ; written in terms of the number of unknowns $N = n^d$ one has $\kappa = \Theta(N^{2/d})$ [34]. That is the scaling quoted in Chapter 9.

5.7.2 The nine-point stencil and other extensions

The five-point molecule ignores the four diagonal neighbours. Including them with weights chosen to match as many Taylor terms as possible produces the standard nine-point stencil [33, 29]

$$(\Delta_h^{(9)} U)_{i,j} = \frac{1}{6h^2} \begin{bmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{bmatrix} U_{i,j}, \quad (5.85)$$

i.e.

$$\begin{aligned} 6h^2 (\Delta_h^{(9)} U)_{i,j} &= 4(U_{i-1,j} + U_{i+1,j} + U_{i,j-1} + U_{i,j+1}) \\ &\quad + (U_{i-1,j-1} + U_{i+1,j-1} + U_{i-1,j+1} + U_{i+1,j+1}) - 20U_{i,j}. \end{aligned} \quad (5.86)$$

A two-dimensional Taylor expansion now yields

$$\Delta_h^{(9)} u = \Delta u + \frac{h^2}{12} \Delta^2 u + O(h^4) = \Delta u + \frac{h^2}{12} (u_{xxxx} + 2u_{xxyy} + u_{yyyy}) + O(h^4). \quad (5.87)$$

Compared with (5.80), the mixed term $2u_{xxyy}$ has appeared: the $O(h^2)$ error is a multiple of $\Delta^2 u$ and is therefore more nearly isotropic. For a plain Laplacian the scheme is still second order. For Poisson's equation $-\Delta u = f$ one can exploit (5.87) by moving the error onto the right-hand side (the *Mehrstellen* or compact scheme). Since $\Delta^2 u = \Delta(\Delta u) = -\Delta f$,

$$-\Delta_h^{(9)} U = f + \frac{h^2}{12} \Delta_h f \quad (5.88)$$

is consistent of order $O(h^4)$ while still using only the nine-point neighbourhood. The left-hand matrix has $s = 9$; the right-hand side is a sparse averaging of f rather than a point sample.

Two other “nine-point” molecules should not be confused with (5.85). Taking the Kronecker sum of the one-dimensional fourth-order stencil (5.69) produces a *cross* of radius two: the center, two steps in each axis direction, and no diagonals. That operator is also nine-point and $O(h^4)$, but it reaches further toward the boundary and is a different sparse pattern. A naive equal-weight average of all eight neighbours is nine-point as well, yet it does not match Δu even to order $O(h^2)$ with a stable positive stencil; the weights in (5.85) are not optional.

On a periodic grid the nine-point operator is still a polynomial in the commuting shifts of Section 5.9.1. Writing $S_x = S \otimes I$ and $S_y = I \otimes S$, the positive Laplacian associated with (5.85) is

$$A^{(9)} = \frac{1}{6h^2} \left(20I - 4(S_x + S_x^\dagger + S_y + S_y^\dagger) - (S_x S_y + S_x S_y^\dagger + S_x^\dagger S_y + S_x^\dagger S_y^\dagger) \right). \quad (5.89)$$

It remains block-circulant with circulant blocks, hence still diagonal in the two-dimensional Fourier basis. The only change for a block encoding is that the LCU now has nine unitary terms rather than five, and the subnormalization grows from $8/h^2$ to $20/(3h^2)$. Dirichlet or Neumann nine-point matrices are obtained from (5.89) by deleting or reflecting the wrap-around terms, exactly as in the five-point case; near a corner as many as five of the nine weights may move into $\mathbf{f}$.

5.7.3 Three dimensions, anisotropy, and variable coefficients

In three dimensions the same construction gives the seven-point stencil ($s = 7$): center weight $-6/h^2$ for Δ_h , plus the six axial neighbours. The full $3 \times 3 \times 3$ neighbourhood is a 27-point molecule; compact $O(h^4)$ schemes analogous to Mehrstellen live in that family. The matrix is the Kronecker sum of three copies of T , and $\kappa = \Theta(h^{-2})$ is unchanged.

If the mesh spacings h_x and h_y differ, the five-point operator becomes

$$(\Delta_h U)_{i,j} = \frac{U_{i-1,j} - 2U_{i,j} + U_{i+1,j}}{h_x^2} + \frac{U_{i,j-1} - 2U_{i,j} + U_{i,j+1}}{h_y^2}. \quad (5.90)$$

The Kronecker form is $I \otimes T_x + T_y \otimes I$ with two different tridiagonal factors. A periodic LCU uses the same shifts with $\alpha = 4/h_x^2 + 4/h_y^2$ [34].

Variable coefficients destroy translation invariance. For $-(a(x)u')' = f$ with $a > 0$ the conservative three-point stencil evaluates a at half-points,

$$\frac{a_{j+1/2}(U_j - U_{j+1}) + a_{j-1/2}(U_j - U_{j-1})}{h^2} = f(x_j), \quad (5.91)$$

where $a_{j\pm 1/2} = a(x_j \pm h/2)$. The matrix remains tridiagonal and SPD, but it is no longer Toeplitz, let alone circulant: each row has its own weights. The DFT/DST diagonalization of Sections 5.9.2 and 5.10 is then unavailable, while a sparse-access block encoding still applies, because $s = 3$ is unchanged. The same distinction—constant-coefficient stencils are polynomials in a shift, variable-coefficient stencils are merely sparse—is the one Li uses when comparing QFT filters with generic QLSP encodings [34].

None of these extensions changes the type of the algebraic problem. They change s , the bandwidth, whether A commutes with a cyclic permutation, and the constant in $\|A\| = \Theta(h^{-2})$. A quantum algorithm that takes a block encoding of A as input inherits those parameters and nothing else.

5.8 Neumann conditions: assembling the matrix

Dirichlet data remove boundary nodes from the unknown vector. Neumann data do the opposite: the boundary value is unknown, and the prescribed flux must be turned into a modified row. Consider

$$-u''(x) = f(x), \quad 0 < x < 1, \quad u'(0) = \sigma_0, \quad u'(1) = \sigma_1. \quad (5.92)$$

Take N equal intervals, $h = 1/N$, and nodes $x_j = jh$ for $j = 0, \dots, N$. The unknown is now the full grid function $\mathbf{u} = (U_0, U_1, \dots, U_N)^\top \in \mathbb{R}^{N+1}$. Three standard devices, following [29, 33], produce three different matrices.

First-order one-sided difference. The forward difference $(U_1 - U_0)/h = \sigma_0$ is a legitimate equation, but it is only $O(h)$ accurate. Inserted into a globally second-order scheme it pulls the whole method down to first order.

Ghost node. Introduce a fictitious value U_{-1} outside the domain and replace the left Neumann condition by the centered formula

$$\frac{U_1 - U_{-1}}{2h} = \sigma_0 \quad \implies \quad U_{-1} = U_1 - 2h\sigma_0. \quad (5.93)$$

The second-difference equation at x_0 for $-u'' = f$ still involves U_{-1} :

$$\frac{-U_{-1} + 2U_0 - U_1}{h^2} = f(x_0). \quad (5.94)$$

Substituting the ghost value eliminates U_{-1} and yields

$$\frac{2}{h^2}(U_0 - U_1) = f(x_0) - \frac{2\sigma_0}{h}. \quad (5.95)$$

At the right end the same argument with $U_{N+1} = U_{N-1} + 2h\sigma_1$ produces

$$\frac{2}{h^2}(U_N - U_{N-1}) = f(x_N) + \frac{2\sigma_1}{h}. \quad (5.96)$$

The Neumann data have entered the right-hand side, scaled by h^{-1} , just as Dirichlet data entered scaled by h^{-2} . Equations (5.95) and (5.96) are consistent but not symmetric: the first row is $(2, -2, 0, \dots)/h^2$ while the neighbouring interior row still has -1 in the $(1, 0)$ slot. Multiplying the two boundary equations by $1/2$ restores symmetry without changing the solution. That factor is the half-cell at each end (a finite-volume reading of the same stencil). The working matrix and right-hand side are then

$$A^{\text{Neu}} = \frac{1}{h^2} \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}, \quad \mathbf{f}^{\text{Neu}} = \begin{pmatrix} \frac{1}{2}f(x_0) - \sigma_0/h \\ f(x_1) \\ \vdots \\ f(x_{N-1}) \\ \frac{1}{2}f(x_N) + \sigma_1/h \end{pmatrix}. \quad (5.97)$$

This is the graph Laplacian of the path on $N + 1$ vertices, scaled by h^{-2} . Every row sums to zero, so $A^{\text{Neu}}\mathbf{1} = 0$. Homogeneous Neumann data $\sigma_0 = \sigma_1 = 0$ leave A^{Neu} unchanged and only sample f (with half-weights at the endpoints).

One-sided second-order difference. Using (5.67) at x_0 gives

$$\frac{-3U_0 + 4U_1 - U_2}{2h} = \sigma_0, \quad (5.98)$$

and the symmetric formula at x_N . The truncation error is $O(h^2)$, but the first row of the system is no longer a second-difference stencil. The matrix ceases to be symmetric, which matters both for classical SPD solvers and for quantum algorithms that prefer a Hermitian block encoding.

5.8.1 Robin conditions

Dirichlet and Neumann are the two extremes of one linear boundary condition, the Robin (or impedance) condition of Chapter 2:

$$(\alpha u + \beta \partial_n u)|_{x=0,1} = g, \quad (5.99)$$

with $(\alpha, \beta) \neq (0, 0)$. At $x = 0$ the outward normal is $-\partial_x$, and at $x = 1$ it is $+\partial_x$. Dirichlet is $\beta = 0$ (the nodal value is data and is removed from the unknown vector). Neumann is $\alpha = 0$ (the previous subsection). A mixed Dirichlet–Neumann problem is Robin with $(\alpha, \beta) = (1, 0)$ at one end and $(0, 1)$ at the other; there is no separate algebraic type.

Assume $\beta_0 \neq 0$ at the left end, and write $\gamma_0 = \alpha_0/\beta_0$ together with g_0 for the left datum. The boundary law is then $\alpha_0 U_0 - \beta_0 u'(0) = g_0$, so $u'(0) = (\alpha_0 U_0 - g_0)/\beta_0$. The centred ghost formula $(U_1 - U_{-1})/(2h) = u'(0)$ gives $U_{-1} = U_1 - (2h\alpha_0/\beta_0)U_0 + (2h/\beta_0)g_0$. Substitute this into the second-difference equation $(-U_{-1} + 2U_0 - U_1)/h^2 = f(x_0)$:

$$\left(\frac{2}{h^2} + \frac{2\alpha_0}{\beta_0 h}\right)U_0 - \frac{2}{h^2}U_1 = f(x_0) + \frac{2g_0}{\beta_0 h}. \quad (5.100)$$

The same $1/2$ scaling that restored symmetry for Neumann produces the first row

$$\left(\frac{1}{h^2} + \frac{\gamma_0}{h}\right)U_0 - \frac{1}{h^2}U_1 = \frac{1}{2}f(x_0) + \frac{g_0}{\beta_0 h}. \quad (5.101)$$

An analogous calculation at $x = 1$ with $\gamma_1 = \alpha_1/\beta_1$ adds γ_1/h to the last diagonal. The assembled operator is therefore the Neumann matrix plus a boundary mass:

$$A^{\text{Rob}} = A^{\text{Neu}} + \frac{1}{h} \text{diag}(\gamma_0, 0, \dots, 0, \gamma_1). \quad (5.102)$$

If $\gamma_0, \gamma_1 \geq 0$ and at least one is positive, then A^{Rob} is SPD: the extra diagonal term kills the kernel of A^{Neu} , so the problem is a QLSP instance with no projection onto $\mathbf{1}^\perp$. If $\beta_0 = 0$, the left node is Dirichlet and is deleted as in (5.76); the remaining matrix has a 2 in its first interior diagonal slot and a Robin (or Neumann) modification at the other end.

The eigenfunctions are no longer a plain DCT. Separation of variables gives $\cos(\mu x + \phi)$ with a transcendental equation for μ that depends on (γ_0, γ_1) ; the DFT/DCT dictionary of Section 5.10 therefore applies only in the Dirichlet, Neumann, and periodic limits, while a generic Robin operator is a sparse Hermitian matrix to be block-encoded directly.

5.8.2 Pure Neumann data and the kernel

If every boundary is Neumann, A^{Neu} in (5.97) is singular [29, 30]. The constant vector $\mathbf{1} = (1, \dots, 1)^\top$ is a null vector because each row sums to zero (discrete conservation of flux). Algebraically $(A^{\text{Neu}})^{-1}$ does not exist, so the problem is not a QLSP instance until the kernel is removed.

Physically, if $\hat{\mathbf{u}}$ is a solution then so is $\hat{\mathbf{u}} + c\mathbf{1}$. Existence requires compatibility: summing all rows of $A^{\text{Neu}}\mathbf{u} = \mathbf{f}^{\text{Neu}}$ gives $0 = \sum_j f(x_j)h - (\sigma_1 - \sigma_0)$, the trapezoid form of $\int_0^1 f = \sigma_1 - \sigma_0$, itself the integral of $-u'' = f$. Uniqueness is restored by a side constraint: fix the mean of $\mathbf{u}$, or pin a single nodal value. Equivalently one inverts A^{Neu} on $\mathbf{1}^\perp$. On the quantum side the same kernel is a zero eigenvalue that must be projected out before an inverse filter is applied.

The spectrum of (5.97) is explicit. The discrete cosine transform (DCT-I) diagonalizes the path Laplacian:

$$(v^{(k)})_j = \cos\left(\frac{k\pi j}{N}\right), \quad \lambda_k(A^{\text{Neu}}) = \frac{4}{h^2} \sin^2\left(\frac{k\pi}{2N}\right), \quad k = 0, \dots, N. \quad (5.103)$$

Thus $\lambda_0 = 0$ and $\lambda_{\max} = \Theta(h^{-2})$. Restricted to $\mathbf{1}^\perp$ one still has $\kappa = \Theta(h^{-2})$. The cosine modes have vanishing discrete derivative at both ends, which is why they are the Neumann eigenfunctions. They are the even part of a Fourier series on a reflected grid; Section 5.10 makes that precise.

5.9 Periodic conditions, circulants, and the DFT

On a circle (or a torus) every node is interior. Identify $x = 0$ with $x = 1$, take N points $x_j = jh$ with $h = 1/N$, and set $U_N \equiv U_0$. The unknown is $\mathbf{u} = (U_0, \dots, U_{N-1})^\top$. The three-point stencil (5.73) now wraps around: the neighbour of U_0 on the left is U_{N-1} , and the neighbour of U_{N-1} on the right is U_0 . Writing the system out,

$$A^{\text{Per}} = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 & -1 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & -1 \\ 0 & \cdots & & -1 & 2 & -1 \\ -1 & 0 & \cdots & 0 & -1 & 2 \end{pmatrix}, \quad \mathbf{f}_j^{\text{Per}} = f(x_j). \quad (5.104)$$

The two corner entries $-1/h^2$ are the whole difference between (5.104) and the Dirichlet Toeplitz matrix (5.76). Those corners are not a nuisance: they are the statement that the grid is a cycle rather than a path, and they make the matrix *circulant*.

5.9.1 The cyclic shift as a permutation

Index the canonical basis by $j = 0, \dots, N-1$ and let S be the permutation matrix of the N -cycle $(0 \ 1 \ \cdots \ N-1)$,

$$S|j\rangle = |j+1 \bmod N\rangle. \quad (5.105)$$

As an array,

$$S = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & & \ddots & 0 & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}. \quad (5.106)$$

Thus S has ones on the subdiagonal and in the top-right corner. It is a permutation matrix, hence unitary: $S^{-1} = S^\top = S^\dagger$ and $S^N = I$. The inverse is the opposite cycle, $S^\dagger|j\rangle = |j-1 \bmod N\rangle$. On grid functions, $(S\mathbf{u})_j = U_{j-1}$ and $(S^\dagger\mathbf{u})_j = U_{j+1}$ (indices modulo N).

The powers $\{I, S, S^2, \dots, S^{N-1}\}$ are the regular representation of the cyclic group $\mathbb{Z}/N\mathbb{Z}$. Every circulant matrix is a polynomial in this single permutation [34]:

$$C = c_0 I + c_1 S + c_2 S^2 + \dots + c_{N-1} S^{N-1}. \quad (5.107)$$

Equivalently, each row of C is a right cyclic shift of the row above. The periodic Laplacian is the three-term case

$$A^{\text{Per}} = \frac{2I - S - S^\dagger}{h^2} = \frac{2I - S - S^{N-1}}{h^2}, \quad (5.108)$$

since $S^\dagger = S^{N-1}$. The same language writes the periodic first differences of Section 5.6,

$$D_+ = \frac{S^\dagger - I}{h}, \quad D_- = \frac{I - S}{h}, \quad D_0 = \frac{S^\dagger - S}{2h}, \quad (5.109)$$

and recovers $A^{\text{Per}} = -D_+ D_- = -D_- D_+$. Because S is unitary, each of $D_\pm$ is an exact linear combination of two unitaries; that is the LCU block encoding of Chapter 6.

On a periodic square the same construction is a Kronecker sum. With $S_x = S \otimes I$ and $S_y = I \otimes S$,

$$A^{\text{Per}, 2\text{D}} = \frac{4I - S_x - S_x^\dagger - S_y - S_y^\dagger}{h^2} = I \otimes A^{\text{Per}} + A^{\text{Per}} \otimes I. \quad (5.110)$$

The matrix is block-circulant with circulant blocks: translating the whole grid by one cell in x or in y leaves it invariant. The nine-point operator (5.89) is the same statement with four extra diagonal shifts $S_x S_y$ and its adjoints; it is still a polynomial in $\{S_x, S_y\}$ and still Fourier-diagonal.

5.9.2 Diagonalization by the DFT

A permutation that is a single N -cycle is diagonalized by the characters of $\mathbb{Z}/N\mathbb{Z}$, i.e. by the discrete Fourier transform. Let $\omega = e^{2\pi i/N}$ and let F be the unitary Fourier matrix

$$F_{jk} = \frac{1}{\sqrt{N}} \omega^{jk} = \frac{1}{\sqrt{N}} e^{2\pi i jk/N}, \quad j, k = 0, \dots, N-1. \quad (5.111)$$

The columns $\varphi^{(k)}$ of F are the Fourier modes $(\varphi^{(k)})_j = N^{-1/2} \omega^{jk}$. Applying the shift,

$$S \varphi^{(k)} = \omega^{-k} \varphi^{(k)}, \quad (5.112)$$

so

$$S = F \Omega F^\dagger, \quad \Omega = \text{diag}(1, \omega^{-1}, \dots, \omega^{-(N-1)}). \quad (5.113)$$

Every circulant, being a polynomial in S , is diagonal in the same basis:

$$C = F \text{diag}(\lambda_0, \dots, \lambda_{N-1}) F^\dagger, \quad \lambda_k = \sum_{m=0}^{N-1} c_m \omega^{-mk}. \quad (5.114)$$

The vector (λ_k) is itself a (non-normalized) DFT of the first column of C . Matrix-vector multiplication by a circulant is therefore one DFT, a pointwise product, and an inverse DFT; inversion of a circulant is the same with $\lambda_k \mapsto 1/\lambda_k$ wherever $\lambda_k \neq 0$.

For the Laplacian (5.108) the symbol is $2 - \omega^{-k} - \omega^k = 2 - 2 \cos(2\pi k/N)$, hence

$$\lambda_k(A^{\text{Per}}) = \frac{4}{h^2} \sin^2\left(\frac{\pi k}{N}\right), \quad k = 0, \dots, N-1. \quad (5.115)$$

In particular $\lambda_0 = 0$ with eigenvector $\mathbf{1}$, so Poisson on a torus is solvable only for mean-zero f , and the inverse is defined on the orthogonal complement—the same kernel as in the Neumann problem, now visible as the $k = 0$ Fourier mode. The two-dimensional operator (5.110) is diagonalized by $F \otimes F$, with $\lambda_{k,\ell} = (\lambda_k + \lambda_\ell)$ a sum of one-dimensional symbols.

Any matrix function of a circulant collapses to a scalar function of the symbol: $f(A^{\text{Per}}) = F f(\Lambda) F^\dagger$. Periodic heat, wave, and Poisson problems are the same DFT with $f(z) = e^{-\kappa t z}$, $\cos(ct\sqrt{z})$, or $1/z$ (on $k \neq 0$). That is the discrete form of the functional calculus of Chapter 11, and it is why a quantum Fourier transform followed by an arithmetic filter is a complete periodic solver [34].

5.10 Which transform for which boundary

The three trigonometric transforms are not three unrelated diagonalizations. They are the Fourier transform on a cycle, restricted to functions with a chosen symmetry [33, 34].

Start with N periodic nodes and the DFT (5.111). Even and odd extensions to a reflected grid produce Neumann and Dirichlet spectra as follows.

- *Periodic*. Translation invariance: $U_{j+N} = U_j$. The symmetry group is the cycle generated by S . The eigenbasis is the full set of Fourier modes $\varphi^{(k)}$, and A^{Per} is circulant.
- *Dirichlet*. Odd extension across the endpoints: place zeros at $j = 0$ and $j = n + 1$ and reflect with a sign change. The resulting $2(n + 1)$ -periodic grid function is an odd combination of Fourier modes, which reduces to the sine vector in (5.77). The wrap-around corners of S are deleted, and the matrix is Toeplitz but not circulant. The DST is the DFT of that odd extension.
- *Neumann*. Even extension: the ghost rule $U_{-1} = U_1$ is reflection without a sign change. The $2N$ -periodic even extension is a cosine combination of Fourier modes, which reduces to (5.103). Algebraically one replaces the wrap-around corners of S by a reflection permutation rather than deleting them; the matrix is the path Laplacian (5.97), diagonalized by DCT-I.
- *Robin*. The extra diagonal mass γ/h in (5.102) breaks even extension, so the eigenfunctions are not a DCT; one block-encodes the sparse matrix directly.

In short: the cyclic permutation S generates the circulant algebra and the DFT; Dirichlet is S with wrap-around removed (odd/sine); Neumann is S with wrap-around replaced by a reflection (even/cosine). Li’s circuit remark is the same picture in gates: start from the periodic shift and correct the boundary-crossing entries [34].

The correspondence is summarized in Table 5.1. On a tensor-product grid the two-dimensional operators are Kronecker sums, so the eigenbases are tensor products of the one-dimensional trigonometric modes (and the fast transform is a product of 1D transforms). That is the classical reason a spectral Poisson solver is an FFT, DST, or DCT; it is also the reason a QFT-based elliptic filter exists for periodic data, with DST/DCT variants for Dirichlet and Neumann boxes (Chapter 12). Generic Robin data fall outside that dictionary: the matrix remains sparse and Hermitian, but its eigenbasis is no longer trigonometric.

5.11 Finite elements, briefly

Finite differences are pointwise. Finite elements start from the weak form of a second-order elliptic problem [29]. For $-\nabla \cdot (A \nabla u) + cu = f$ with homogeneous Dirichlet data one seeks $u \in H_0^1(\Omega)$

boundary	matrix	group action	transform
periodic	circulant (5.104)	cycle S	DFT
Dirichlet	Toeplitz (5.76)	odd extension	DST
Neumann	path Laplacian (5.97)	even extension	DCT-I
Robin	(5.102)	A^{Neu} plus boundary mass	not a DCT

Table 5.1: Second-difference matrices, the symmetry that produces them from the cyclic shift, and the trigonometric transform that diagonalizes them. In every case $\lambda_k = (4/h^2) \sin^2 \theta_k$ for a grid-dependent angle θ_k , so $\kappa = \Theta(h^{-2})$ on the complement of any kernel.

such that

$$a(u, v) = \langle f, v \rangle \quad \text{for all } v \in H_0^1(\Omega), \quad (5.116)$$

where $a(u, v) = \int_{\Omega} (A \nabla u \cdot \nabla v + cuv)$ and Dirichlet data have already been built into the space (the trace of H_0^1 vanishes on $\partial\Omega$). Neumann data, by contrast, appear naturally as a boundary integral after integration by parts; they modify the linear functional on the right-hand side rather than the trial space.

A Galerkin subspace $V_h \subset H_0^1(\Omega)$ with basis $\{\phi_j\}_{j=1}^N$ produces the stiffness matrix and load vector

$$K_{ij} = a(\phi_j, \phi_i), \quad F_i = \langle f, \phi_i \rangle, \quad K\mathbf{U} = \mathbf{F}. \quad (5.117)$$

Compact support of the basis keeps K sparse and, for the usual elliptic coercivity, SPD. The algebraic problem is again a QLSP instance. Two structural differences from (5.76) matter later: the load vector samples f against basis functions rather than at nodes, and a mass matrix $M_{ij} = \int \phi_j \phi_i$ appears as soon as the problem is time-dependent, yielding $M\mathbf{u}' = K\mathbf{u}$ rather than $\mathbf{u}' = A\mathbf{u}$. Mass lumping replaces M by a diagonal matrix and returns one to the finite-difference cost model [34]. For piecewise linears the basis functions are the hats (5.42), and the entries of K , M , and F are integrals of the kind computed in Section 5.5.

Finite-volume and discontinuous Galerkin schemes assemble different sparse matrices (cell averages and numerical fluxes; locally invertible mass blocks) [29]. They are the natural discretizations of conservation laws in Chapter 13, but they still produce either $A\mathbf{u} = \mathbf{f}$ or $\mathbf{u}' = A\mathbf{u}$.

5.12 Method of lines and time stepping

Discretize space and leave time continuous. For the heat equation $u_t = \kappa u_{xx}$ on $(0, 1)$ with homogeneous Dirichlet data, the interior vector $\mathbf{u}(t)$ of Section 5.7 satisfies

$$\mathbf{u}'(t) = -\kappa A\mathbf{u}(t) + \mathbf{g}(t), \quad \mathbf{u}(0) = \mathbf{u}_0, \quad (5.118)$$

where A is the SPD matrix (5.76) and $\mathbf{g}$ collects any inhomogeneous boundary or forcing terms that have been moved to the right-hand side, exactly as α/h^2 was. This is a linear ODE. If A is constant, Duhamel's formula [29] gives

$$\mathbf{u}(t) = e^{-\kappa t A} \mathbf{u}_0 + \int_0^t e^{-\kappa(t-s)A} \mathbf{g}(s) ds. \quad (5.119)$$

The homogeneous piece is the matrix semigroup $e^{-\kappa t A}$.

A time step of size $k = \Delta t$ produces a fully discrete scheme. Forward Euler is explicit; the von Neumann calculation of Section 5.13.3 forces the parabolic restriction $k \leq h^2/(2\kappa)$. Backward

Euler and the trapezoidal (Crank–Nicolson) rule are implicit. Averaging the spatial operator at t_n and t_{n+1} yields

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{k} = \frac{\kappa}{2}(-A\mathbf{u}^n - A\mathbf{u}^{n+1}) \quad (5.120)$$

in the homogeneous case, or, with the diffusion number $r = \kappa k / (2h^2)$ written out on the three-point stencil [29],

$$-r U_{j-1}^{n+1} + (1 + 2r)U_j^{n+1} - r U_{j+1}^{n+1} = r U_{j-1}^n + (1 - 2r)U_j^n + r U_{j+1}^n. \quad (5.121)$$

Each step is a tridiagonal linear system $(I + \frac{\kappa k}{2}A)\mathbf{u}^{n+1} = \mathbf{b}^n$. The left-hand matrix is SPD with condition number still $\Theta(h^{-2})$ for fixed k , so every implicit step is a QLSP instance (Chapter 9, reused in Chapter 14). Truncation and stability of Crank–Nicolson are computed in Section 5.13: the scheme is $O(k^2 + h^2)$ and unconditionally stable, but not L -stable.

5.12.1 Absolute stability

The scalar model of (5.118) is $y' = \lambda y$. A linear multistep step of width k ,

$$\sum_{j=0}^m \alpha_j y_{n+j} = k \sum_{j=0}^m \beta_j f(t_{n+j}, y_{n+j}), \quad \alpha_m = 1, \quad (5.122)$$

applied to this equation produces a linear recurrence. Its characteristic polynomial is $\rho(\zeta) - z\sigma(\zeta)$, where $z = k\lambda$, $\rho(\zeta) = \sum_{j=0}^m \alpha_j \zeta^j$, and $\sigma(\zeta) = \sum_{j=0}^m \beta_j \zeta^j$. The value z lies in the region of absolute stability when every root satisfies $|\zeta| \leq 1$ and every root of modulus 1 is simple [32].

Forward Euler, $y_{n+1} = y_n + k f_n$, amplifies by $1+z$, so the region is the disk $|1+z| \leq 1$. Backward Euler, $y_{n+1} = y_n + k f_{n+1}$, amplifies by $1/(1-z)$. The trapezoidal rule, which is Crank–Nicolson for (5.118), amplifies by $(1+z/2)/(1-z/2)$. For every z with negative real part the last two factors have modulus at most 1. Along the negative real axis, as $|z| \rightarrow \infty$, the trapezoidal factor tends to -1 . A mode with $k|\lambda| \gg 1$ therefore decays slowly and changes sign at each step, which is the failure of L -stability recorded above.

The root condition is the same restriction at $z = 0$. Consistency forces $\rho(1) = 0$, and the condition asks that every root of ρ satisfy $|\zeta| \leq 1$, with every root on the unit circle simple. The explicit two-step formula $y_{n+1} = 3y_n - 2y_{n-1}$ has $\rho(\zeta) = (\zeta - 1)(\zeta - 2)$. On $y' = 0$ the root 2 produces a factor 2^n , so the recurrence grows for every step size. For a consistent multistep method, the root condition, stability under perturbations of the starting values, and convergence are equivalent.

The eigenvalues of $-\kappa A$ in (5.118) are real and negative, and the largest has size $\Theta(\kappa/h^2)$. Forward Euler keeps $z = k\lambda$ inside $|1+z| \leq 1$ only for $k = O(h^2)$, which is the parabolic restriction $k \leq h^2/(2\kappa)$ of Section 5.13.3. Backward Euler and Crank–Nicolson contain the negative axis, so the step size is free of that restriction, and each step is the linear system already written, with condition number still $\Theta(h^{-2})$.

The wave equation $u_{tt} = c^2 u_{xx}$, reduced to a first-order system in $(\mathbf{u}, \mathbf{u}_t)$ or factored through a discrete gradient, produces a generator that is skew-Hermitian (or becomes Hermitian after a Schrödingerization / first-order dilation). The natural matrix function is then e^{-itH} , i.e. Hamiltonian simulation, rather than a dissipative semigroup [34]. First-order transport $u_t + au_x = 0$ is the scalar model of that hyperbolic class. Upwind differences respect the characteristic direction and are stable under the CFL condition $|a|k/h \leq 1$; centered second-order schemes such as Lax–Wendroff raise the order but, by Godunov’s theorem, cannot stay monotone while remaining

linear [29]. Those design issues belong to Chapter 13. The only fact needed here is that the spatial operator is again a sparse matrix, typically non-normal, acting on a grid function whose boundary treatment follows the same Dirichlet/Neumann/periodic playbook.

5.13 Consistency, stability, and convergence

Assembling $A\mathbf{u} = \mathbf{f}$ does not by itself guarantee that $\mathbf{u}$ is close to the PDE solution. The missing ingredients are consistency of the stencil, stability of the inverse (or of the time stepper), and the implication from those two to convergence [33, 29]. A quantum algorithm inherits the discretization error unchanged: it replaces a classical solve of the same linear algebra, at a different cost in κ and s .

5.13.1 The abstract error equation

Let L_h be a linear discrete operator on a mesh of size h , with $L_h U = f_h$ the scheme and u a smooth exact solution of the PDE. The *local truncation error* is the residual of the exact solution in the discrete equations,

$$\tau_h := L_h(u|_{\text{grid}}) - f_h. \quad (5.123)$$

The scheme is *consistent* of order p if $\|\tau_h\| = O(h^p)$ as $h \rightarrow 0$ (and, for time stepping, $O(k^q + h^p)$ jointly). The *global error* $e := U - u|_{\text{grid}}$ then satisfies

$$L_h e = -\tau_h \quad (5.124)$$

whenever $L_h U = f_h$. Hence

$$\|e\| \leq \|L_h^{-1}\| \|\tau_h\|, \quad (5.125)$$

provided L_h is invertible. Consistency controls $\|\tau_h\|$; *stability* is the statement that $\|L_h^{-1}\|$ remains bounded as $h \rightarrow 0$, uniformly in the mesh. For linear schemes those two hypotheses are necessary and sufficient for convergence $\|e\| \rightarrow 0$ (Lax–Richtmyer) [33]. The same identity is why a one-sided first-order Neumann row, whose truncation is $O(h)$, pulls a globally second-order interior stencil down to $O(h)$: $\|\tau_h\|$ is already only $O(h)$, and no bound on $\|A^{-1}\|$ can recover an order that the residual does not have.

The two-norm of A^{-1} is $1/\lambda_{\min}(A)$. That quantity being $O(1)$ (Dirichlet Poisson on a fixed domain) is ℓ^2 stability of the elliptic scheme. It is not to be confused with the condition number $\kappa(A) = \lambda_{\max}/\lambda_{\min} = \Theta(h^{-2})$, which governs round-off and the query cost of a quantum linear solver, not the discretization error.

5.13.2 Elliptic schemes: from τ_h to a global $O(h^2)$

For the Dirichlet problem (5.71) one has $L_h = A$ from (5.76) and, by (5.65),

$$\tau_j = (Au|_h - f)_j = -\frac{h^2}{12} u^{(4)}(x_j) + O(h^4). \quad (5.126)$$

Thus $\|\tau\|_{\infty} = O(h^2)$ as soon as $u \in C^4$. The DST spectrum (5.77) gives $\lambda_{\min}(A) = \Theta(1)$, so (5.125) already yields $\|e\|_2 = O(h^2)$. A maximum-norm bound follows from the discrete maximum principle, which is the statement that A is inverse-positive: $Av \geq 0$ (componentwise) implies $v \geq 0$ for interior grid functions vanishing at the Dirichlet ends [33].

To convert this into an explicit constant, let $\psi_j = \frac{1}{2}x_j(1 - x_j)$. Then $\psi'' = -1$ and $\psi^{(4)} = 0$, so the three-point stencil is exact on ψ and $(A\psi)_j = 1$. If $|\tau_j| \leq M$, the comparison functions $M\psi \pm e$ satisfy

$$A(M\psi - e) = M\mathbf{1} + \tau \geq 0, \quad A(M\psi + e) = M\mathbf{1} - \tau \geq 0, \quad (5.127)$$

because $Ae = -\tau$ by (5.124). Inverse-positivity yields $M\psi \pm e \geq 0$ and

$$\max_j |e_j| \leq M \max_j \psi_j = \frac{M}{8}. \quad (5.128)$$

With $M = (h^2/12)\|u^{(4)}\|_\infty + O(h^4)$ one obtains

$$\max_j |U_j - u(x_j)| \leq \frac{h^2}{96} \|u^{(4)}\|_\infty + O(h^4). \quad (5.129)$$

The five-point scheme on a square is the same argument in two dimensions: $\tau = O(h^2)$ by (5.80), A remains inverse-positive, and a quadratic comparison function (e.g. $\psi = \frac{1}{2}x(1 - x) + \frac{1}{2}y(1 - y)$) gives $\max |e| = O(h^2)$. The nine-point Mehrstellen modification (5.88) replaces τ by $O(h^4)$, so the same stability produces a global $O(h^4)$ provided the boundary closures match that order.

5.13.3 Time stepping: amplification and CFL

For a linear constant-coefficient evolution, substitute a Fourier mode $U_j^n = \hat{U}^n e^{i\xi x_j}$ into the fully discrete scheme. Each step becomes multiplication by a scalar *amplification factor* $g(\xi)$, and stability in ℓ^2 on a periodic (or frozen-coefficient) grid is $|g(\xi)| \leq 1$ for all frequencies ξ [33, 29]. The eigenvalues of the three-point Laplacian on such a mode are $\lambda(\xi) = (4/h^2) \sin^2(\xi h/2)$, the same symbol as (5.115).

Forward Euler, $(U^{n+1} - U^n)/k = -\kappa A U^n$, has

$$g(\xi) = 1 - \kappa k \lambda(\xi) = 1 - 4\rho \sin^2\left(\frac{\xi h}{2}\right), \quad \rho := \frac{\kappa k}{h^2}. \quad (5.130)$$

The most restrictive value is $\sin^2 = 1$, so $|g| \leq 1$ for all ξ if and only if $4\rho \leq 2$, i.e.

$$k \leq \frac{h^2}{2\kappa}. \quad (5.131)$$

This is a CFL condition of parabolic type: time steps must refine as h^2 , not as h . Dirichlet sine modes replace ξh by $k\pi/(n+1)$ and give the same restriction, attained in the highest frequency.

Crank–Nicolson averages A at the two time levels. On a Fourier mode

$$g(\xi) = \frac{1 - 2\rho \sin^2(\xi h/2)}{1 + 2\rho \sin^2(\xi h/2)}. \quad (5.132)$$

For $\rho > 0$ the numerator is strictly smaller in absolute value than the denominator, so $|g(\xi)| < 1$ with no restriction on k : the scheme is unconditionally stable. As $\rho \rightarrow \infty$, $g \rightarrow -1$, so large steps are stable but oscillatory (the scheme is A -stable, not L -stable) [29].

The local truncation error of Crank–Nicolson is read from Taylor’s theorem in t about the midpoint $t_{n+1/2}$, together with the spatial expansion (5.65). The left-hand time difference equals $u_t + (k^2/24)u_{ttt} + O(k^4)$ at that point; the averaged second difference equals $\kappa u_{xx} + (\kappa h^2/12)u_{xxx} + O(h^4 + k^2)$. Using $u_t = \kappa u_{xx}$ the leading terms cancel and

$$\tau = O(k^2 + h^2). \quad (5.133)$$

Lax–Richtmyer with the uniform bound $|g| \leq 1$ therefore gives a global error $O(k^2 + h^2)$ on a fixed time interval. Forward Euler is only $O(k + h^2)$, and only inside the regime (5.131).

For the linear advection equation $u_t + au_x = 0$ the same Fourier substitution on a three-point spatial stencil produces a CFL condition of hyperbolic type, $|a|k/h \leq 1$: the numerical domain of dependence must contain the characteristic $x - at = \text{const}$ [29, 33]. Upwind is first-order and monotone inside that regime; Lax–Wendroff is second-order and, by Godunov’s theorem, cannot be both linear and monotonicity-preserving. Those schemes are developed in Chapter 13.

5.13.4 Galerkin error

Finite elements replace truncation of a stencil by orthogonality in the energy inner product. If $a(\cdot, \cdot)$ is continuous with constant α and coercive with constant β on $H_0^1(\Omega)$, Lax–Milgram gives a unique weak solution. The Galerkin condition $a(u - u_h, v_h) = 0$ for all $v_h \in V_h$ then implies, for arbitrary $v_h \in V_h$,

$$\begin{aligned} \beta \|u - u_h\|_{H^1}^2 &\leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) \\ &\leq \alpha \|u - u_h\|_{H^1} \|u - v_h\|_{H^1}, \end{aligned} \quad (5.134)$$

hence Céa’s lemma $\|u - u_h\|_{H^1} \leq (\alpha/\beta) \inf_{v_h \in V_h} \|u - v_h\|_{H^1}$ [29]. On shape-regular meshes of size h , the infimum for piecewise linears is $O(h)|u|_{H^2}$, so the energy error is $O(h)$ and, under elliptic regularity, the L^2 error is $O(h^2)$ by a duality argument. The algebraic system $K\mathbf{U} = \mathbf{F}$ is still a QLSP instance; Céa’s constant plays the role of $\|L_h^{-1}\|$ in (5.125).

5.13.5 Discretization error in the quantum cost model

An end-to-end error budget therefore has at least three terms [34]: the global discretization error just estimated, the QLSP or simulation error ε of Chapters 9 and 8, and the sampling error of the observable. Matching $O(h^p)$ to a target accuracy ε_{tot} forces $h = \Theta(\varepsilon_{\text{tot}}^{1/p})$ and, for second-order elliptic stencils,

$$\kappa(A) = \Theta(h^{-2}) = \Theta(\varepsilon_{\text{tot}}^{-2/p}). \quad (5.135)$$

Refining the mesh to reduce discretization error inflates the condition number that a quantum linear solver pays for. That coupling is the reason Li tracks mesh-dependent cost rather than treating A as an abstract matrix [34], and it is why first-order factorization (reducing the spectral scale from h^{-2} to h^{-1}) is a discretization-level decision, not a circuit identity.

5.14 Conditioning, sparsity, and what a quantum algorithm sees

End-to-end cost is not the cost of inverting A alone [34]. After the constructions of this chapter, the quantities that reappear in Chapters 6 and 9 are already determined.

- *Dimension.* N is the number of retained unknowns. On a d -dimensional mesh, $N = \Theta(h^{-d})$, stored in $\lceil \log_2 N \rceil$ qubits.
- *Sparsity.* The stencil width s is the number of nodes in the molecule of Section 5.6.6: $s = 3$ for the three-point second derivative, $s = 5$ for the five-point Laplacian or the one-dimensional fourth-order stencil, $s = 9$ for the nine-point Laplacian, and $s = 2d + 1$ for the standard d -dimensional second-order Laplacian. FEM stiffness matrices have s set by the element degree and mesh topology, still $O(1)$ per row for local bases.

- *Norm and subnormalization.* $\|A\| = \Theta(h^{-2})$ for a second-order elliptic stencil. A block encoding represents A/α with $\alpha \geq \|A\|$; the periodic LCU of Section 5.9.1 takes $\alpha = 4/h^2$ in one dimension and $\alpha = 8/h^2$ in two dimensions [34].
- *Fourier structure.* Periodic matrices are polynomials in the cyclic permutation S and are diagonalized by the DFT (Section 5.9.2). Dirichlet and Neumann matrices are the odd and even restrictions of the same picture (DST/DCT, Section 5.10). A QFT (or a DST/DCT circuit) is therefore a spectral encoding of the discrete Laplacian, not an extra algorithm. Generic Robin matrices lose that trigonometric eigenbasis (Section 5.8.1).
- *Condition number.* $\kappa(A) = \Theta(h^{-2}) = \Theta(N^{2/d})$ for second-order elliptic and implicit parabolic matrices on a domain of fixed size. This, not N itself, is the dominant parameter in HHL and QSVT inversion.
- *Right-hand side.* Dirichlet and Neumann data live in $\mathbf{f}$ (or in $\mathbf{g}(t)$). Preparing a normalized $|b\rangle \propto \mathbf{f}$ is a state-preparation cost, cheap for smooth or structured data and expensive for arbitrary classical vectors.
- *Kernel.* Pure Neumann or periodic Poisson operators are singular until the constant mode is removed. An inverse filter cannot be applied on a spectral window that contains 0.
- *Discretization error.* Consistency plus stability give $\|e\| = O(h^p)$ independently of how $A\mathbf{u} = \mathbf{f}$ is solved (Section 5.13). Matching that order to a target accuracy inflates $\kappa = \Theta(h^{-2})$.
- *Readout.* The quantum output is a normalized state, not the N nodal values. Observables, inner products, or derived PDE quantities are the intended measurements; writing down every entry of $\mathbf{u}$ destroys the asymptotic accounting.

The next chapter turns the sparse stencil—shifts $S, S^\dagger$ in the periodic case, or a sparse-access oracle once boundaries have been corrected—into a block encoding. From that point on the pipeline is linear algebra.

Part II

Fault-Tolerant Quantum Linear Algebra and Primitives

Chapter 6

Block encoding and qubitization

A block encoding represents a (normalized) matrix A/α as a distinguished block of a larger unitary [40, 35, 34]. Once a discretized differential operator is in this form, QSVT, Hamiltonian simulation, and linear-system solvers become compositions of a small set of primitives—the analogue of BLAS/LAPACK for quantum scientific computing.

The chapter has four steps. Section 6.1 defines a block encoding and shows what the subnormalization α costs. Section 6.2 lists the rules for combining block encodings, and Section 6.3 proves the most important one, the linear combination of unitaries. Section 6.4 applies these rules to the stencil matrices of Chapter 5, where the cyclic shift S turns a finite-difference Laplacian into a sum of four or five unitaries. Section 6.5 then shows how a single block-encoding unitary is iterated to produce Chebyshev polynomials of A/α in the same block, which is the starting point of Chapters 7, 8 and 9. The development follows Lin and Wiebe, Chapters 9 and 10 [35], with the finite-difference examples taken from Li [34].

6.1 Block encodings

A quantum circuit is a unitary matrix, and the matrices of Chapter 5 are not. The discrete Laplacian is real symmetric with eigenvalues in $[0, 4d/h^2]$, while a unitary has all of its eigenvalues on the unit circle. A circuit therefore cannot apply A to a state $|b\rangle$, and the operation $|b\rangle \mapsto A|b\rangle/\|A|b\rangle\|$ is not a gate. It can still be obtained by embedding A , suitably scaled, into a larger unitary and keeping only the part of the output in which an extra register has returned to $|0\rangle$.

Let $A \in \mathbb{C}^{N \times N}$ with $N = 2^n$. A unitary U acting on $a + n$ qubits is an (α, a, ε) -block-encoding of A if

$$\|A - \alpha(|0^a\rangle\langle 0^a| \otimes I_N) U (|0^a\rangle\langle 0^a| \otimes I_N)\| \leq \varepsilon. \quad (6.1)$$

The a qubits are the *ancilla* register. The projector $\Pi = |0^a\rangle\langle 0^a| \otimes I_N$ singles out the top-left block of U , and for $\varepsilon = 0$ the definition says

$$U = \begin{pmatrix} A/\alpha & * \\ * & * \end{pmatrix} \quad (6.2)$$

in a basis ordered with the ancilla state $|0^a\rangle$ first. The scalar $\alpha > 0$ is the *subnormalization*. The entries $*$ are whatever unitarity forces; they are never used. A unitary V is an $(1, 0, 0)$ -block-encoding of itself, since no ancilla is needed. The weaker data (α, a) with $\varepsilon = 0$ are written $\text{BE}_{\alpha,a}(A)$.

The first constraint is $\|A\| \leq \alpha + \varepsilon$, because a submatrix of a unitary has norm at most one. Hence $\alpha \geq \|A\|$ up to the error, and the ratio $\alpha/\|A\| \geq 1$ measures how much the encoding wastes. Existence is never the obstacle. If $A = A^\dagger$ and $\|A\| \leq 1$, then A and $\sqrt{I - A^2}$ commute and

$$U = \begin{pmatrix} A & \sqrt{I - A^2} \\ \sqrt{I - A^2} & -A \end{pmatrix} \quad (6.3)$$

satisfies $U^2 = I$ and $U = U^\dagger$, so it is a unitary $(1, 1, 0)$ -block-encoding. For a general A with $\|A\| \leq 1$ the Halmos dilation has the off-diagonal blocks $\sqrt{I - AA^\dagger}$ and $\sqrt{I - A^\dagger A}$ and $-A^\dagger$ in the corner. The difficulty is that (6.3) requires the matrix $\sqrt{I - A^2}$, which is as hard to apply as A itself. The real input model is therefore not the dilation but a *circuit* for U of small depth, and the rest of this chapter constructs such circuits for the matrices of Chapter 5.

To see what the block does, apply U to $|0^a\rangle|b\rangle$. Splitting the output according to Π gives

$$U|0^a\rangle|b\rangle = |0^a\rangle \frac{A|b\rangle}{\alpha} + |\perp\rangle, \quad (\langle 0^a| \otimes I)|\perp\rangle = 0. \quad (6.4)$$

Measuring the ancilla and keeping the outcome 0^a therefore succeeds with probability

$$p = \frac{\|A|b\rangle\|^2}{\alpha^2}, \quad (6.5)$$

and the system register then holds $A|b\rangle/\|A|b\rangle\|$. Amplitude amplification raises p to a constant at the price of $O(\alpha/\|A|b\rangle\|)$ further uses of U . This is the mechanism by which α enters every complexity bound of later chapters. For the periodic Laplacian of Section 6.4 with $\alpha = \|A\|$, a state $|b\rangle$ concentrated at the Nyquist frequency is accepted with probability close to one. A smooth state is not: the lowest nonzero Fourier mode has $A|b\rangle = \lambda_1|b\rangle$ with $\lambda_1/\alpha = \sin^2(\pi/N) \approx (\pi h)^2$, so $p \approx (\pi h)^4$ and the amplified cost is $O(h^{-2}) = O(\kappa)$. Smooth states live at the bottom of the spectrum, and the block A/α is small there by construction.

Finally, a block encoding can be checked entry by entry: $(\langle 0^a|i\rangle U(|0^a\rangle|j\rangle)) = A_{ij}/\alpha$. This is how the circuits below are verified, and it is also how an approximate encoding is understood, since (6.1) bounds the norm of the difference between A and the realized block.

6.2 An algebra of block encodings

Block encodings compose in the way matrices do, but the normalizations multiply or add and the ancilla registers accumulate. Three rules suffice for this report; the proofs are one-line projections [35, 34].

Adjoint. If U_A is an (α, a, ε) -block-encoding of A , then $U_A^\dagger$ is an (α, a, ε) -block-encoding of $A^\dagger$, because $\Pi U_A^\dagger \Pi = (\Pi U_A \Pi)^\dagger$.

Tensor product. If U_A and U_B are exact block encodings with data (α, a) and (β, b) , acting on different system registers, then $U_A \otimes U_B$ is an exact $(\alpha\beta, a + b)$ -block-encoding of $A \otimes B$ once the ancilla registers are grouped together. This is used together with the LCU rule below, because a Kronecker sum $A \otimes I + I \otimes B$ is a sum of two tensor products, each with a block-encoded identity factor.

Product. Let U_A be an $(\alpha, a, \varepsilon_A)$ -block-encoding of A and U_B a $(\beta, b, \varepsilon_B)$ -block-encoding of B . Order the registers as $\mathcal{H}_a \otimes \mathcal{H}_b \otimes \mathcal{H}_{\text{sys}}$ and let $\tilde{U}_A$ act as U_A on (a, sys) and as the identity on b , and let $\tilde{U}_B$ act as U_B on (b, sys) and as the identity on a . The register a stays in $|0^a\rangle$ while $\tilde{U}_B$ acts, and b is untouched while $\tilde{U}_A$ acts, so in the exact case

$$(\langle 0^{a+b}| \otimes I) \tilde{U}_A \tilde{U}_B (|0^{a+b}\rangle \otimes I) = \frac{A}{\alpha} \frac{B}{\beta} = \frac{AB}{\alpha\beta}. \quad (6.6)$$

Thus $\tilde{U}_A \tilde{U}_B$ is an $(\alpha\beta, a+b, \varepsilon_{AB})$ -block-encoding of AB . Write $A = \alpha\tilde{A} + E_A$ and $B = \beta\tilde{B} + E_B$ with $\|\tilde{A}\|, \|\tilde{B}\| \leq 1$ the realized blocks and $\|E_A\| \leq \varepsilon_A$, $\|E_B\| \leq \varepsilon_B$. Then $AB - \alpha\beta\tilde{A}\tilde{B} = \alpha\tilde{A}E_B + \beta E_A\tilde{B} + E_A E_B$, so

$$\varepsilon_{AB} \leq \alpha\varepsilon_B + \beta\varepsilon_A + \varepsilon_A\varepsilon_B. \quad (6.7)$$

The normalization factors multiply, which is the quantum counterpart of the fact that $\|AB\| \leq \|A\|\|B\|$. The product rule applied k times gives a block encoding of A^k with normalization α^k , i.e. the normalized block is $(A/\alpha)^k$, but it spends ka ancilla qubits. Section 6.5 removes that growth: qubitization produces polynomials of A/α in the same a ancillas.

Sum. Sums of block encodings are handled by the linear combination of unitaries of the next section, where the normalizations add instead of multiplying.

6.3 Linear combination of unitaries

Suppose a matrix is given as a weighted sum of unitaries,

$$T = \sum_{i=0}^{K-1} \alpha_i U_i, \quad \alpha_i > 0, \quad (6.8)$$

where a complex or negative coefficient has been absorbed into the unitary (if $\alpha_i = |\alpha_i|e^{i\phi_i}$, use $e^{i\phi_i}U_i$; the matrix $-S$ is unitary whenever S is). Pad K to 2^a and let $\|\alpha\|_1 = \sum_i \alpha_i$. Define the two circuits

$$\text{PREP } |0^a\rangle = \sum_{i=0}^{K-1} \sqrt{\frac{\alpha_i}{\|\alpha\|_1}} |i\rangle, \quad \text{SELECT} = \sum_{i=0}^{K-1} |i\rangle\langle i| \otimes U_i + \sum_{i \geq K} |i\rangle\langle i| \otimes I, \quad (6.9)$$

where the second sum in SELECT only keeps the operator unitary on the padded labels, which have zero amplitude after PREP. The register of a qubits is the *selector*: PREP writes the coefficients in superposition, and SELECT applies the unitary indexed by the selector value to the system register. Then

$$W = (\text{PREP}^\dagger \otimes I) \text{SELECT} (\text{PREP} \otimes I) \quad (6.10)$$

is an exact $(\|\alpha\|_1, a, 0)$ -block-encoding of T . The proof is a direct projection: since $(\langle 0^a | \text{PREP}^\dagger) = \sum_i \sqrt{\alpha_i / \|\alpha\|_1} \langle i |$ and SELECT acts as U_i on the system when the selector is $|i\rangle$,

$$(\langle 0^a | \otimes I) W (|0^a\rangle \otimes I) = \sum_i \sqrt{\frac{\alpha_i}{\|\alpha\|_1}} \sqrt{\frac{\alpha_i}{\|\alpha\|_1}} U_i = \frac{T}{\|\alpha\|_1}. \quad (6.11)$$

Figure 6.1 shows the circuit. Equation (6.11) is the block-matrix identity

$$\begin{pmatrix} w_0 I & \cdots & w_{K-1} I \end{pmatrix} \begin{pmatrix} U_0 & & \\ & \ddots & \\ & & U_{K-1} \end{pmatrix} \begin{pmatrix} w_0 I \\ \vdots \\ w_{K-1} I \end{pmatrix} = \sum_i w_i^2 U_i, \quad w_i = \sqrt{\alpha_i / \|\alpha\|_1}, \quad (6.12)$$

in which the selector chooses one diagonal block and PREP supplies the weights on both sides.

Because the ancilla outcome 0^a is postselected, W acts on a normalized state $|\psi\rangle$ with success probability $\|T|\psi\rangle\|^2 / \|\alpha\|_1^2$ by (6.5). The coefficient sum $\|\alpha\|_1$ is therefore the normalization of the encoding and not merely a convention. It can be much larger than $\|T\|$ when the terms cancel, for instance when T is a small difference of two nearly equal unitaries, and a decomposition with a large

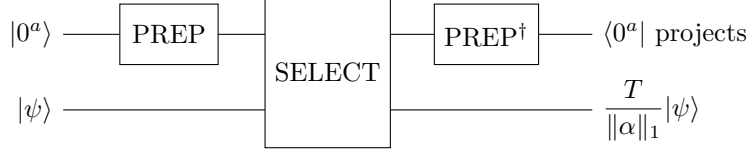


Figure 6.1: The LCU circuit (6.10). The upper wire is the selector register with $a = \lceil \log_2 K \rceil$ qubits and the lower wire is the system. When the selector returns to $|0^a\rangle$ the system holds $T|\psi\rangle/\|\alpha\|_1$.

ℓ^1 norm is expensive even if $\|T\|$ is small. Among splits of the form $\alpha_i = \bar{\beta}_i \gamma_i$, with PREP preparing $\beta/\|\beta\|_2$ on the left and $\gamma/\|\gamma\|_2$ on the right, the normalization is $\|\beta\|_2\|\gamma\|_2 \geq \sum_i |\beta_i||\gamma_i| = \|\alpha\|_1$ by the Cauchy–Schwarz inequality. Equality holds exactly for the symmetric choice $|\beta_i| = |\gamma_i| = \sqrt{|\alpha_i|}$ of (6.9), which is therefore the optimal split. An asymmetric split is still useful when the two sides of a circuit must be different circuits; Section 6.5 meets that case.

The LCU rule extends to sums of block encodings. If U_i is an $(\gamma_i, a', \varepsilon_i)$ -block-encoding of A_i and $A = \sum_i c_i A_i$, then the same circuit, with SELECT built from the U_i and PREP built from the weights $|c_i|\gamma_i$, is a block encoding of A with normalization $\sum_i |c_i|\gamma_i$, with $a + a'$ ancillas and with error $\sum_i |c_i|\varepsilon_i$. Normalizations and errors add. A continuous parameter is handled by quadrature: if $f(A) = \int w(s)g(s, A) ds$ is approximated by $\sum_k w_k g(s_k, A)$ with $g(s_k, A)$ unitary, the result is an LCU with $\|\alpha\|_1 = \sum_k |w_k|$. The Cauchy quadratures of Chapter 11 and the LCHS formula of Section 10 have this form, and $\sum_k |w_k|$ plays the role of a stability constant of the quadrature rule. LCU was introduced for Hamiltonian simulation by Childs and Wiebe [16].

A caution on cost. The selector needs only $\lceil \log_2 K \rceil$ qubits, so a sum of exponentially many terms fits in a small register. This is a saving only when PREP and SELECT have compact circuits: for unstructured coefficients the preparation of the state in (6.9) and the multi-controlled SELECT may themselves cost $O(K)$. The next section is the case where both are trivial, since $K = 2d + 1$ and the unitaries are shifts.

6.4 Sparse-matrix and finite-difference block encodings

Stencil matrices from Sec. 5 are the first concrete encodings used later for Poisson, heat, and wave operators [34]. The unifying fact is that a constant-coefficient stencil is a polynomial in the cyclic shift S of Section 5.9.1, and S is already unitary. Set $N = 2^n$ and $S|j\rangle = |j + 1 \bmod N\rangle$. The circuit for S is modular addition of one, a reversible incrementer on n qubits, and the circuit for $S^\dagger$ is the decrementer. These have gate counts polynomial in $n = \log_2 N$; explicit circuits for the banded circulant and tridiagonal cases are due to Camps, Lin, Van Beeumen and Yang [13].

6.4.1 Periodic stencils as a linear combination of shifts

By (5.108) the periodic three-point matrix is

$$A^{\text{Per}} = \frac{2I - S - S^\dagger}{h^2} = \frac{1}{h^2}(2I + 1 \cdot (-S) + 1 \cdot (-S^\dagger)), \quad (6.13)$$

an LCU with the three unitaries $I, -S, -S^\dagger$ and coefficients $2/h^2, 1/h^2, 1/h^2$. The coefficient sum is $\alpha = 4/h^2$. A selector of two qubits holds the three labels, with the fourth label unused. Take

$$\text{PREP } |0\rangle = \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{2}|1\rangle + \frac{1}{2}|2\rangle, \quad \text{SELECT} = |0\rangle\langle 0| \otimes I + |1\rangle\langle 1| \otimes (-S) + |2\rangle\langle 2| \otimes (-S^\dagger) + |3\rangle\langle 3| \otimes I. \quad (6.14)$$

Equation (6.11) gives the top-left block $\frac{1}{2}I - \frac{1}{4}S - \frac{1}{4}S^\dagger = h^2 A^{\text{Per}}/4$, so the circuit $U_L = (\text{PREP}^\dagger \otimes I) \text{SELECT} (\text{PREP} \otimes I)$ is an exact $(4/h^2, 2, 0)$ -block-encoding of A^{Per} . The circuit uses the incrementer and decremter once each, controlled on the selector, and a constant number of rotations.

This encoding wastes nothing. The eigenvalues of A^{Per} are $\lambda_k = (4/h^2) \sin^2(\pi k/N)$ (Section 5.9.2), so for even N the largest equals $4/h^2$ and $\alpha = \|A^{\text{Per}}\|$ exactly. The normalized matrix A^{Per}/α has Fourier eigenvalues $\sin^2(\pi k/N) \in [0, 1]$. Its smallest value is zero (the constant mode of the periodic problem) and its smallest nonzero value is $\sin^2(\pi/N) = \Theta(h^2)$, which is the condition number $\kappa = \Theta(h^{-2})$ restated in the normalized scale. The unitary U_L is not Hermitian, because $-S \neq -S^\dagger$ makes $\text{SELECT}^\dagger \neq \text{SELECT}$; Section 6.5 therefore uses its general form, with both U_L and $U_L^\dagger$.

On a periodic square the matrix is a Kronecker sum, and with $S_x = S \otimes I$, $S_y = I \otimes S$ the five-point operator is

$$A_{2D}^{\text{Per}} = \frac{4I - S_x - S_x^\dagger - S_y - S_y^\dagger}{h^2}. \quad (6.15)$$

This is an LCU of five unitaries with coefficient sum $\alpha = 8/h^2$. Three selector qubits hold the five labels, and PREP prepares $\frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{8}}(|1\rangle + |2\rangle + |3\rangle + |4\rangle)$. The weights are $\frac{1}{2}$ for the identity and $\frac{1}{8}$ for each of the four shifts. Again $\alpha = \|A_{2D}^{\text{Per}}\|$. For d dimensions the same construction gives

$$A_{dD}^{\text{Per}} = \frac{1}{h^2} \sum_{r=1}^d (2I - S_r - S_r^\dagger), \quad \alpha = \frac{4d}{h^2}, \quad (6.16)$$

with $2d + 1$ unitary terms, $\lceil \log_2(2d + 1) \rceil$ selector qubits, and PREP amplitudes $\sqrt{1/2}$ on the identity and $\sqrt{1/(4d)}$ on each shift. The number of selector qubits grows only with $\log d$, and the grid size N enters only through the cost of the shifts, which is polynomial in $\log N$. For anisotropic spacings $h_x \neq h_y$ the PREP amplitudes are the square roots of the absolute stencil coefficients, and $\alpha = 4/h_x^2 + 4/h_y^2$.

The nine-point operator (5.89) is an LCU of nine unitaries, among them the products $S_x S_y$, and has $\alpha = 20/(3h^2)$ with four selector qubits. Here the encoding is not tight. At the highest frequency $S_x = S_y = -1$ each of the four axis shifts contributes $+4$ and each of the four corner products $S_x S_y = +1$ contributes -1 , so the operator takes the value $(20 + 16 - 4)/(6h^2) = 16/(3h^2) = \|A^{(9)}\|$. The coefficient sum adds the absolute values, $(20 + 16 + 4)/(6h^2)$, so $\alpha/\|A\| = 5/4$. A smaller ratio would require a different decomposition.

An alternative route to the same matrix uses the DFT. Since $A^{\text{Per}} = F^\dagger \Lambda F$ with F the quantum Fourier transform and Λ the diagonal of eigenvalues, a block encoding is the product $F^\dagger \cdot \Lambda \cdot F$ with the diagonal $\Lambda/\|\Lambda\|$ encoded by a controlled rotation whose angle is computed from $\sin^2(\pi k/N)$. By (6.6) the normalization is $\|\Lambda\| = 4/h^2$, the same as the LCU. The shift LCU needs no arithmetic beyond a modular increment, whereas the spectral route needs trigonometric arithmetic but keeps working for any circulant. This is the sense in which Section 5.14 calls a QFT a spectral encoding of the Laplacian.

6.4.2 Boundary conditions

Dirichlet and Neumann matrices are the periodic matrix with the wrap-around terms deleted or reflected. Work on a path with $M = 2^n$ nodes (a size that is not a power of two is padded, for instance with an identity block or with dummy eigenvalues outside the spectral window of interest). Let $\Pi_0 = |0\rangle\langle 0|$, $\Pi_{\text{end}} = |M-1\rangle\langle M-1|$, and $D = I - \Pi_{\text{end}}$. The matrix SD maps $|j\rangle \mapsto |j+1\rangle$ for $j < M-1$ and sends $|M-1\rangle$ to zero, so it is the shift without the wrap-around entry, with ones

on the subdiagonal; its transpose is $DS^\dagger$. The Dirichlet matrix of (5.76) is therefore

$$A^{\text{Dir}} = \frac{2I - SD - DS^\dagger}{h^2}. \quad (6.17)$$

The diagonal projector D has an exact $(1, 1, 0)$ -block-encoding: add one flag qubit, initially $|0\rangle$, and flip it when the node index equals $M-1$, that is $U_D = I \otimes D + X_{\text{flag}} \otimes \Pi_{\text{end}}$. Then $\langle 0|U_D|0\rangle = D$. The product rule (6.6) makes SD and $DS^\dagger$ exact $(1, 1, 0)$ -block-encodings (the shift needs no ancilla), and the LCU of Section 6.3 with the same coefficients 2, 1, 1 gives

$$\alpha = \frac{4}{h^2}, \quad a = 2 + 1, \quad (6.18)$$

two selector qubits and one flag. The boundary correction costs one comparator and one ancilla, and it leaves α unchanged. Since $\|A^{\text{Dir}}\| = (2 + 2 \cos(\pi/(M+1)))/h^2 = (4 - O(h^2))/h^2$, the encoding is again tight up to a factor $1 + O(h^2)$.

The Neumann matrix (5.97) differs from the Dirichlet one by half the diagonal at the two ends, $A^{\text{Neu}} = (2E - SD - DS^\dagger)/h^2$ with $E = I - \frac{1}{2}(\Pi_0 + \Pi_{\text{end}})$. The diagonal matrix E has entries $\frac{1}{2}$ at the end nodes and 1 elsewhere, so a flag rotation with cosine $\frac{1}{2}$, controlled on the node being an endpoint, is an exact $(1, 1, 0)$ -block-encoding of it. The data are again $\alpha = 4/h^2$ and $a = 3$. A Robin condition adds a diagonal term γ/h supported on the end nodes, a fourth LCU term whose coefficient enters α additively. In each case the conclusion is the one stated in Section 5: the boundary changes a constant number of entries, and the block encoding changes by one comparator.

6.4.3 General sparse access and variable coefficients

Constant-coefficient stencils are special in that A is a polynomial in a unitary. A variable-coefficient or finite-element matrix is merely sparse. Assume at most s nonzeros per row and column, with $f(i, \ell)$ the column of the ℓ th nonzero in row i (padded to length s), and assume coherent access to the two oracles

$$O_{\text{loc}} |i\rangle|\ell\rangle|0\rangle = |i\rangle|\ell\rangle|f(i, \ell)\rangle, \quad O_{\text{val}} |i\rangle|\ell\rangle|0\rangle = |i\rangle|\ell\rangle|A_{i, f(i, \ell)}\rangle. \quad (6.19)$$

For finite differences these oracles are reversible arithmetic ($f(i, \ell) \in \{i-1, i, i+1\}$ and a stencil weight); for finite elements they enumerate the elements that touch a basis function. Write $A_{ij} = |A_{ij}|e^{i\theta_{ij}}$ and choose $\alpha \geq \max_i \sum_j |A_{ij}|$ and $\alpha \geq \max_j \sum_i |A_{ij}|$. On an enlarged space with a basis $|i, j\rangle$ of nonzero positions (“edges”) define the isometries

$$\begin{aligned} P|i\rangle &= \frac{1}{\sqrt{\alpha}} \sum_{j: A_{ij} \neq 0} \sqrt{|A_{ij}|} |i, j\rangle + \sqrt{1 - \frac{1}{\alpha} \sum_j |A_{ij}|} |i, \perp\rangle, \\ Q|j\rangle &= \frac{1}{\sqrt{\alpha}} \sum_{i: A_{ij} \neq 0} e^{i\theta_{ij}} \sqrt{|A_{ij}|} |i, j\rangle + \sqrt{1 - \frac{1}{\alpha} \sum_i |A_{ij}|} |\perp, j\rangle, \end{aligned} \quad (6.20)$$

where the garbage states $|i, \perp\rangle$ and $|\perp, j\rangle$ are orthogonal to all edge states and to each other. The only edge shared by $P|i\rangle$ and $Q|j\rangle$ is $|i, j\rangle$, so $\langle i|P^\dagger Q|j\rangle = A_{ij}/\alpha$, that is $P^\dagger Q = A/\alpha$ [34, 35]. Each of P and Q is implemented from a uniform superposition over $\ell = 0, \dots, s-1$, one call each to O_{loc} and O_{val} , a controlled rotation whose amplitude is $\sqrt{|A_{i, f(i, \ell)}|/\alpha}$, a phase rotation $e^{i\theta}$, and an uncomputation of the value register. The crude bound $\alpha = s\|A\|_{\text{max}}$ always suffices, since each row has at most s entries of modulus at most $\|A\|_{\text{max}}$; the sharper choice above is the maximum absolute row or column sum. For the tridiagonal matrix (6.17) the row sums are at most

$4/h^2$, against $s\|A\|_{\max} = 6/h^2$, so the sparse construction reproduces the LCU normalization. For a Hermitian sparse matrix a *Hermitian* block encoding with $n+2$ ancillas and the same normalization is available [35]; this is the form the reflection argument of Section 6.5 uses.

A variable coefficient shows how the two views meet. Take $-(a(x)u')'$ with $a \geq 0$ and the three-point form of (5.91), where the weights are $a_{j\pm 1/2} = a(x_j \pm h/2)$ and $S, S^\dagger$ act on vectors as $(Su)_j = u_{j-1}$, $(S^\dagger u)_j = u_{j+1}$. Then

$$A = \frac{1}{h^2}(\Lambda_+ + \Lambda_- - \Lambda_+ S^\dagger - \Lambda_- S), \quad \Lambda_\pm = \text{diag}(a_{j\pm 1/2}). \quad (6.21)$$

A diagonal matrix with entries bounded by $a_{\max}$ has an $(a_{\max}, 1, \varepsilon)$ -block-encoding: compute the entry $a_{j\pm 1/2}/a_{\max}$ into a register, rotate a flag qubit by the corresponding angle, and uncompute. The product with S keeps the normalization by (6.6). The four terms of (6.21) form an LCU with $\alpha = 4a_{\max}/h^2$. The matrix is not circulant, and the DFT route of the previous section is lost, but the block encoding costs no more than in the constant-coefficient case apart from one coefficient oracle.

6.4.4 Resource summary

Table 6.1 collects the encodings of this section. In every row the cost of one use of U_A is a constant number of shifts and rotations, hence polynomial in $\log N$, while $\alpha = \Theta(h^{-2})$ grows with the mesh. This is the point made in Section 5.14: a cheap sparse query does not by itself give a polylogarithmic solver, because a polynomial of A/α must resolve a spectral window of relative width h^2 .

Operator	terms	subnormalization α	ancillas
1D periodic, (6.13)	3	$4/h^2$	2
2D periodic, five-point, (6.15)	5	$8/h^2$	3
d D periodic, (6.16)	$2d + 1$	$4d/h^2$	$\lceil \log_2(2d + 1) \rceil$
2D periodic, nine-point, (5.89)	9	$20/(3h^2)$	4
1D Dirichlet, (6.17)	3	$4/h^2$	$2 + 1$
1D Neumann	3	$4/h^2$	$2 + 1$
1D variable coefficient, (6.21)	4	$4a_{\max}/h^2$	$2 + 1$

Table 6.1: Block encodings of stencil matrices as a linear combination of shifts. The columns give the number of unitary terms, the normalization α (the coefficient 1-norm), and the ancilla count (selector qubits, plus flag qubits for boundary or coefficient oracles). Only the nine-point row has $\alpha > \|A\|$ beyond an $O(h^2)$ correction.

6.5 Qubitization

Given a block encoding U of A/α we want polynomials of that matrix in a block of a unitary. The product rule does it with one new ancilla register per factor, and simply iterating U does not do it: the top-left block of U^2 is $(A/\alpha)^2$ plus a contribution from the off-diagonal blocks $*$ of (6.2). *Qubitization* is the observation that, after one reflection is inserted between uses of U , the off-diagonal blocks cancel and the powers produce Chebyshev polynomials in the same ancilla register [40, 35]. Below, the normalization is absorbed and A denotes the matrix with $\|A\| \leq 1$, that is A/α in the notation of the previous sections.

6.5.1 Hermitian case: reflection and rotation

Let $A = A^\dagger = \sum_i \lambda_i |v_i\rangle\langle v_i|$ with $|\lambda_i| \leq 1$, and let $U = U^\dagger$ be a Hermitian $(1, m, 0)$ -block-encoding, as in (6.3). Put $\Pi = |0^m\rangle\langle 0^m| \otimes I$ and $Z_\Pi = 2\Pi - I$. By (6.4),

$$U|0^m\rangle|v_i\rangle = \lambda_i|0^m\rangle|v_i\rangle + \sqrt{1 - \lambda_i^2}|\perp_i\rangle, \quad \Pi|\perp_i\rangle = 0, \quad (6.22)$$

for a normalized $|\perp_i\rangle$ (when $|\lambda_i| = 1$ the second term vanishes and the subspace below is one-dimensional, with the statements true trivially). Because $U^2 = I$, applying U to (6.22) gives

$$U|\perp_i\rangle = \sqrt{1 - \lambda_i^2}|0^m\rangle|v_i\rangle - \lambda_i|\perp_i\rangle. \quad (6.23)$$

Hence the two-dimensional space $\mathcal{H}_i = \text{span}\{|0^m\rangle|v_i\rangle, |\perp_i\rangle\}$ is invariant under U , and the whole space is a direct sum of such planes: each eigenvector of A is *qubitized* into its own qubit. In the ordered basis, U and Π are

$$U|_{\mathcal{H}_i} = \begin{pmatrix} \lambda_i & \sqrt{1 - \lambda_i^2} \\ \sqrt{1 - \lambda_i^2} & -\lambda_i \end{pmatrix}, \quad Z_\Pi|_{\mathcal{H}_i} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (6.24)$$

The first matrix is a *reflection*, which has determinant -1 and squares to the identity; this is why U^k just alternates between U and I and cannot produce a polynomial. The product of the two reflections in (6.24) is a rotation. With $\lambda_i = \cos \theta_i$,

$$O := U Z_\Pi|_{\mathcal{H}_i} = \begin{pmatrix} \cos \theta_i & -\sin \theta_i \\ \sin \theta_i & \cos \theta_i \end{pmatrix}, \quad O^k|_{\mathcal{H}_i} = \begin{pmatrix} \cos k\theta_i & -\sin k\theta_i \\ \sin k\theta_i & \cos k\theta_i \end{pmatrix}. \quad (6.25)$$

Geometrically, Z_Π reflects across the axis $|0^m\rangle|v_i\rangle$, U reflects across a line at angle $\theta_i/2$ to it, and two reflections compose to a rotation by twice that angle (Figure 6.2). Since $\cos k\theta = T_k(\cos \theta)$ with T_k the Chebyshev polynomial of the first kind, the top-left block of O^k on $\mathcal{H}_i$ is $T_k(\lambda_i)$, and summing over i ,

$$(|0^m\rangle\langle 0^m| \otimes I) O^k (|0^m\rangle\langle 0^m| \otimes I) = \sum_i T_k(\lambda_i) |v_i\rangle\langle v_i| = T_k(A). \quad (6.26)$$

Thus $O^k = (U Z_\Pi)^k$ is a $(1, m, 0)$ -block-encoding of $T_k(A)$, with the *same* m ancillas as U and with k uses of U and k of Z_Π .

6.5.2 General block encodings

The block encodings of Section 6.4 are not Hermitian, so the reflection argument must be redone with U and $U^\dagger$. Let $U \in \text{BE}_{1,m}(A)$ for a general $A = \sum_i \sigma_i |w_i\rangle\langle w_i|$ with singular values $\sigma_i \in [0, 1]$, written in the singular value decomposition. By (6.4) applied to U and to $U^\dagger$ there are normalized states $|\perp_i\rangle$ and $|\perp'_i\rangle$ in the kernel of Π with

$$U|0^m\rangle|v_i\rangle = \sigma_i|0^m\rangle|w_i\rangle + \sqrt{1 - \sigma_i^2}|\perp'_i\rangle, \quad U^\dagger|0^m\rangle|w_i\rangle = \sigma_i|0^m\rangle|v_i\rangle + \sqrt{1 - \sigma_i^2}|\perp_i\rangle. \quad (6.27)$$

Then U maps $\mathcal{H}_i = \text{span}\{|0^m\rangle|v_i\rangle, |\perp_i\rangle\}$ to $\mathcal{H}'_i = \text{span}\{|0^m\rangle|w_i\rangle, |\perp'_i\rangle\}$, and $U^\dagger$ maps $\mathcal{H}'_i$ back to $\mathcal{H}_i$. In these bases both are the reflection matrix of (6.24) with $\lambda_i \rightarrow \sigma_i$, and $Z_\Pi = \text{diag}(1, -1)$ on both planes. The product $U^\dagger Z_\Pi U Z_\Pi$ preserves $\mathcal{H}_i$ and acts there as the square of the rotation in (6.25), and repeating it k times gives, with $\sigma_i = \cos \theta_i$,

$$(U^\dagger Z_\Pi U Z_\Pi)^k|_{\mathcal{H}_i} = \begin{pmatrix} \cos 2k\theta_i & -\sin 2k\theta_i \\ \sin 2k\theta_i & \cos 2k\theta_i \end{pmatrix}. \quad (6.28)$$

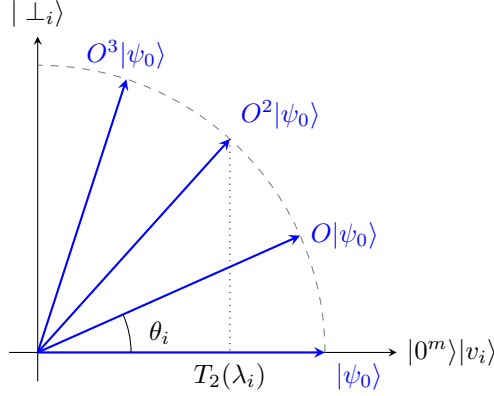


Figure 6.2: The plane $\mathcal{H}_i$ for an eigenvalue $\lambda_i = \cos \theta_i$ of A , with $|\psi_0\rangle = |0^m\rangle|v_i\rangle$. Each use of $O = UZ_\Pi$ rotates the state by θ_i . The component along $|\psi_0\rangle$ after k steps is $\cos k\theta_i = T_k(\lambda_i)$, shown for $k = 2$.

Collecting the planes,

$$(U^\dagger Z_\Pi U Z_\Pi)^k \in \text{BE}_{1,m}(T_{2k}^{\text{SV}}(A)), \quad U Z_\Pi (U^\dagger Z_\Pi U Z_\Pi)^k \in \text{BE}_{1,m}(T_{2k+1}^{\text{SV}}(A)), \quad (6.29)$$

where, for a polynomial p of definite parity, the singular value transformation is defined by [35, 23]

$$p^{\text{SV}}(A) = \begin{cases} \sum_i p(\sigma_i) |w_i\rangle\langle v_i|, & p \text{ odd}, \\ \sum_i p(\sigma_i) |v_i\rangle\langle v_i|, & p \text{ even}. \end{cases} \quad (6.30)$$

The case T_{2k} uses $2k$ queries to U and $U^\dagger$, and T_{2k+1} uses $2k + 1$. If $A = A^\dagger$ the singular value transformation reduces to an eigenvalue transformation: T_k has the parity of k , so $T_k(|\lambda|) \text{sgn}(\lambda)^k = T_k(\lambda)$ and $T_k^{\text{SV}}(A) = T_k(A)$ for every k . For the Hermitian reflection of the previous subsection (6.26) is recovered. The reflection $Z_\Pi = (2|0^m\rangle\langle 0^m| - I_m) \otimes I_n$ is a phase flip on all states of the ancilla register except $|0^m\rangle$, a multi-controlled Z gate on m qubits. It costs $O(m)$ gates and one working qubit. Qubitization therefore adds nothing to the ancilla count of U except that qubit, in contrast with the product rule, whose ancilla count grows linearly in the number of factors.

6.5.3 Example: the periodic Laplacian

Take U_L of Section 6.4.1 with normalization $\alpha = 4/h^2$, so that the normalized matrix is $\tilde{A} = h^2 A^{\text{Per}}/4$ with Fourier eigenvectors $|f_k\rangle$ and eigenvalues $\sin^2(\pi k/N) \in [0, 1]$. The singular values of $\tilde{A}$ coincide with these eigenvalues, since $\tilde{A}$ is positive semidefinite, and each Fourier mode spans its own plane $\mathcal{H}_k$ with the angle

$$\cos \theta_k = \sin^2\left(\frac{\pi k}{N}\right), \quad \lambda_k = \frac{4}{h^2} \cos \theta_k. \quad (6.31)$$

By (6.29) the circuit $(U_L^\dagger Z_\Pi U_L Z_\Pi)^j$ acts on $|f_k\rangle$ with block $\cos(2j\theta_k) = T_{2j}(\sin^2(\pi k/N))$. The quantum walk $U_L^\dagger Z_\Pi U_L Z_\Pi$ has eigenvalues $e^{\pm 2i\theta_k}$, so the discrete Laplacian spectrum has been turned into *phases*; this is what Low and Chuang exploit for Hamiltonian simulation [40]. The two ends of the spectrum behave differently. The Nyquist mode has $\cos \theta = 1$, $\theta = 0$, and is a fixed

point of the rotation. All smooth modes sit at $\cos \theta_k \approx (\pi k h)^2 \approx 0$, that is $\theta_k = \pi/2 - O(h^2)$, within $O(h^2)$ of the same angle. Because $T_j(x) = \cos(j \arccos x)$ varies by about $j \, dx$ near $x = 0$, resolving $\tilde{A}$ -eigenvalues of size $\sin^2(\pi/N) \approx (\pi h)^2$ requires degree $j = \Omega(h^{-2}) = \Omega(\kappa)$. This is the quantitative form of the statement that a QSVT inverse filter for the Poisson matrix needs $\tilde{O}(\kappa)$ queries (Section 9.4).

6.5.4 Why qubitization is the generic structure

It may seem a coincidence that an arbitrary block-encoding unitary splits into 2×2 blocks. It is not. The cosine-sine (CS) decomposition states that any unitary $V \in \mathbb{C}^{(p+q) \times (p+q)}$ with $q \geq p$ can be written

$$V = \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix} \begin{pmatrix} C & S' & 0 \\ S' & -C & 0 \\ 0 & 0 & I_{q-p} \end{pmatrix} \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix}^\dagger, \quad (6.32)$$

with $W_1, V_1 \in \mathbb{C}^{p \times p}$ and $W_2, V_2 \in \mathbb{C}^{q \times q}$ unitary, and $C = \text{diag}(\cos \theta_i)$, $S' = \text{diag}(\sin \theta_i)$ nonnegative diagonal [35]. The top-left block of V is $W_1 C V_1^\dagger$, which is the singular value decomposition of the encoded matrix, with singular values $\cos \theta_i$. The middle factor, after a permutation, is a direct sum of the 2×2 reflections in (6.24), and Z_Π is a direct sum of $\text{diag}(1, -1)$. Because Z_Π is diagonal in the ancilla register it commutes with the outer unitary factors $\text{diag}(W_1, W_2)$ and $\text{diag}(V_1, V_2)$, which is why the construction does not need to know them. This is also why Z_Π cannot be replaced by a Pauli X or Y on the ancilla.

6.5.5 Chebyshev series and what comes next

Qubitization yields one Chebyshev polynomial per use, $T_{2k}(A)$ in a block of O^k . A polynomial of A is a combination of such blocks, so the LCU rule of Section 6.3 applies. For an even polynomial $f(A) = \sum_{k=0}^{K-1} c_k T_{2k}(A)$ with $K = 2^{a'}$, take $\text{SELECT} = \sum_k |k\rangle\langle k| \otimes U_{2k}$ with $U_{2k} \in \text{BE}_{1,m}(T_{2k}(A))$. Implemented naively, SELECT would need $\sum_k 2k = O(K^2)$ queries. Writing $k = (k_{a'-1} \cdots k_0)_2$ in binary and applying the powers $(U^\dagger Z_\Pi U Z_\Pi)^{2^j}$ controlled on bit k_j reuses the blocks and needs $2K - 2$ queries in total, the cost of the single term T_{2K-2} [35]. With PREP built from the coefficients c_k the result is a $(\|c\|_1, m + a')$ -block-encoding of $f(A)$. Two limitations explain why the report moves on to QSVT in Chapter 7: the normalization is the Chebyshev ℓ^1 norm $\|c\|_1$ and not one, and the selector needs $a' = O(\log K)$ extra qubits. Replacing Z_Π by a phase $e^{i\varphi_j Z_\Pi}$ at each step, with the angles φ_j chosen by quantum signal processing, produces an *arbitrary* polynomial of degree K of definite parity with the normalization one and no selector register, in K queries [23].

The quantum-walk eigenvalues $e^{\pm 2i\theta_i}$ from (6.28) also have two uses that appear in the next chapters. Phase estimation applied to $W = U^\dagger Z_\Pi U Z_\Pi$ recovers θ_i , hence $\lambda_i = \alpha \cos \theta_i$, which is the eigenvalue oracle that HHL needs without any Hamiltonian simulation (Section 9.3). The Jacobi–Anger expansion of e^{-itx} in Chebyshev polynomials turns the qubitized walk into a Hamiltonian simulation of e^{-itH} with $O(\alpha t + \log(1/\epsilon)/\log \log(1/\epsilon))$ steps [40]; this is the subject of Chapter 8. The pipeline of the report is now visible: Section 6.4 converts a PDE stencil into (α, a) ; qubitization converts (α, a) into $T_k(A/\alpha)$; Chapter 7 converts those into any polynomial; and Chapters 8 and 9 use the polynomials for e^{-itA} and A^{-1} .

Chapter 7

Quantum signal processing and singular value transformation

Quantum signal processing (QSP) implements a polynomial of a scalar signal by a product of single-qubit rotations. Quantum singular value transformation (QSVT) lifts that construction to a block-encoded matrix, realizing a polynomial $p(A)$ (or a polynomial of the singular values of a general A) [23, 35, 34].

This chapter develops the *tool*. Three matrix functions used later are mentioned only as targets, not as PDE solvers:

- inverse filters $p(x) \approx 1/x$ (QLSP, Sec. 9.4);
- exponential filters $p(x) \approx e^{-tx}$ (heat semigroup);
- oscillatory propagators $p(x) \approx e^{-itx}$ (Hamiltonian simulation).

On Hermitian (or normal) spectra these polynomials are real-line surrogates of the Cauchy integral that defines $f(A)$ in Sec. 11. Linear combination of Hamiltonian simulation (LCHS) and variational linear solvers (VQLS) are *not* treated here; they belong to Secs. 10 and 16.

The argument has three layers. In Chapter 6 a block encoding U and the reflection Z_Π split the space into planes, one for each singular value $\sigma_i = \cos \theta_i$ of the encoded matrix, and the product UZ_Π acts on every plane as the same 2×2 rotation by θ_i . Section 7.1 studies that rotation as a one-qubit problem: if free phases are inserted between d uses of the rotation, which polynomials of the signal σ_i can appear in the matrix, and how are the phases found? Section 7.2 puts the planes back, so that one set of phases acts on all singular values at once, and gives the circuit and its cost. Section 7.3 collects the polynomials that the later chapters ask for, with their degrees.

7.1 Quantum signal processing

7.1.1 The signal rotation and the phase sequence

Write a signal $x \in [-1, 1]$ as $x = \cos \theta$ with $\theta \in [0, \pi]$ and let

$$O(x) = \begin{pmatrix} x & -\sqrt{1-x^2} \\ \sqrt{1-x^2} & x \end{pmatrix} = e^{-i\theta Y} \quad (7.1)$$

be the *signal rotation*, with Y the Pauli matrix. This is the matrix O of (6.25) with x in place of λ_i or σ_i , and (6.26) says that its power $O(x)^k = O(\cos k\theta)$ has the Chebyshev polynomial $T_k(x)$ in

the top-left entry. The signal x is given and cannot be changed. What can be chosen is a phase gate $e^{i\varphi Z} = \text{diag}(e^{i\varphi}, e^{-i\varphi})$ between the uses of $O(x)$. A *QSP sequence* of length d with phases $\Phi = (\varphi_0, \dots, \varphi_d) \in \mathbb{R}^{d+1}$ is

$$U_\Phi(x) = e^{i\varphi_0 Z} \prod_{j=1}^d (O(x) e^{i\varphi_j Z}) = e^{i\varphi_0 Z} O(x) e^{i\varphi_1 Z} O(x) \dots O(x) e^{i\varphi_d Z}, \quad (7.2)$$

and it uses the signal d times.

With all phases zero the sequence is $O(x)^d$ and its top-left entry is $T_d(x)$. The phases turn this single polynomial into a family. For $d = 2$ a direct multiplication gives the top-left entry $e^{i(\varphi_0+\varphi_2)}(2\cos\varphi_1 x^2 - e^{-i\varphi_1})$, which equals $T_2(x) = 2x^2 - 1$ for $\varphi_1 = 0 = \varphi_0 + \varphi_2$ and is the constant $i e^{i(\varphi_0+\varphi_2)}$ for $\varphi_1 = \pi/2$. Two questions follow. Which polynomials arise, and for a wanted polynomial, which phases produce it? The first has a complete answer. Write $s = \sqrt{1-x^2}$, let $\bar{P}$ denote the polynomial with complex-conjugated coefficients, and call a polynomial even or odd according to its parity 0 or 1.

Structure theorem [23, 35]. A matrix-valued function $U(x)$ equals $U_\Phi(x)$ for some $\Phi \in \mathbb{R}^{d+1}$ if and only if

$$U(x) = \begin{pmatrix} P(x) & -\bar{Q}(x)s \\ Q(x)s & \bar{P}(x) \end{pmatrix} \quad (7.3)$$

for polynomials $P, Q \in \mathbb{C}[x]$ with

1. $\deg P \leq d$ and $\deg Q \leq d-1$;
2. P has parity $d \bmod 2$ and Q has parity $(d-1) \bmod 2$;
3. $|P(x)|^2 + (1-x^2)|Q(x)|^2 = 1$ for all $x \in [-1, 1]$.

Necessity. For $d = 0$ the matrix is $e^{i\varphi_0 Z}$, which has the form (7.3) with $P = e^{i\varphi_0}$ and $Q = 0$. Suppose the product of the first $j-1$ signal rotations has the form (7.3) with the pair (P_{j-1}, Q_{j-1}) . Multiplying from the right by $O(x)e^{i\varphi_j Z}$ gives the same form with

$$P_j = e^{i\varphi_j}(xP_{j-1} - (1-x^2)\bar{Q}_{j-1}), \quad Q_j = e^{i\varphi_j}(xQ_{j-1} + \bar{P}_{j-1}). \quad (7.4)$$

If P_{j-1} and Q_{j-1} have degrees at most $j-1$ and $j-2$ and parities $j-1$ and $j-2$, then the right-hand sides have degrees at most j and $j-1$ and parities j and $j-1$: multiplication by x and conjugation-and-swap each flip the parity, and the factor $1-x^2$ preserves it. Condition (iii) says that the first column of the unitary $U_\Phi(x)$ has norm one.

Sufficiency follows by stripping one factor. For $d = 0$ the conditions force $Q = 0$ and $|P| = 1$, so $P = e^{i\varphi_0}$. Let $d \geq 1$ and let p_d and q_{d-1} be the coefficients of x^d in P and of x^{d-1} in Q . The coefficient of x^{2d} in (iii) gives $|p_d| = |q_{d-1}|$. If both vanish take $\varphi_d = 0$; otherwise choose φ_d with $e^{2i\varphi_d} = p_d/\bar{q}_{d-1}$, a number of modulus one. With $(P_1, Q_1) = e^{-i\varphi_d}(P, Q)$ define

$$P' = xP_1 + (1-x^2)\bar{Q}_1, \quad Q' = xQ_1 - \bar{P}_1. \quad (7.5)$$

This is the inverse of (7.4), so $U = U' O(x) e^{i\varphi_d Z}$, where U' is built from (P', Q') . The choice of φ_d cancels the coefficient of x^{d+1} in P' and of x^d in Q' , and the parities then give $\deg P' \leq d-1$ and $\deg Q' \leq d-2$ with the right parities. Condition (iii) holds for (P', Q') because $U' = U(O(x)e^{i\varphi_d Z})^{-1}$ is again unitary with determinant one, which forces the form (7.3). By induction $U' = U_\Phi$ with $\Phi' = (\varphi_0, \dots, \varphi_{d-1})$.

Real targets. Applications specify a real polynomial F and read out the real part of P . If $F \in \mathbb{R}[x]$ has degree at most d , parity $d \bmod 2$, and $|F| \leq 1$ on $[-1, 1]$, then there are phases with $\operatorname{Re} P_\Phi = F$ [23]. To see this, note that $1 - F^2$ is even and nonnegative on $[-1, 1]$, so $1 - F(x)^2 = r(x^2)$ with $r \geq 0$ on $[0, 1]$ and $\deg r \leq d$. A polynomial that is nonnegative on $[0, 1]$ is a sum $r(y) = a(y)^2 + y(1 - y)b(y)^2$ for even degree and $r(y) = ya(y)^2 + (1 - y)b(y)^2$ for odd degree with real a, b (Lukács). Substituting $y = x^2$ gives real polynomials A, B with

$$1 - F(x)^2 = A(x)^2 + (1 - x^2)B(x)^2, \quad (7.6)$$

where A has degree at most d and parity $d \bmod 2$, and B has degree at most $d - 1$ and parity $(d - 1) \bmod 2$. Then $P = F + iA$ and $Q = B$ satisfy (i)–(iii) since $|P|^2 = F^2 + A^2$, and the structure theorem supplies Φ . The imaginary part A is a *completion*: it is not prescribed, and different completions give different phase sequences for the same F .

7.1.2 Phase factors in practice

Two conventions are in common use. The one above, with the real rotation $O(x)$, matches the qubitization of Chapter 6. Much of the algorithmic literature [35, 54] uses instead the X -rotation $W(x) = e^{i \arccos(x)X}$ and reads out $\operatorname{Im}[U_\Phi^W]_{11}$. Since $O(x) = e^{i\pi Z/4} W(x) e^{-i\pi Z/4}$, the two sequences differ only at the ends. If a W -sequence Φ^W has $\operatorname{Im}[U_{\Phi^W}^W]_{11} = F$, then $\Phi = (\varphi_0^W - \pi/2, \varphi_1^W, \dots, \varphi_d^W)$ is an O -sequence with $U_\Phi = e^{-i\pi Z/4} U_{\Phi^W}^W e^{-i\pi Z/4}$, whose top-left entry is $-i[U_{\Phi^W}^W]_{11}$, so that $\operatorname{Re} P_\Phi = F$. A third, circuit-level, convention appears in Section 7.2.2. Table 7.1 collects them.

	signal operator	target	phases
O (this report)	$O(x) = e^{-i \arccos(x)Y}$	$\operatorname{Re}[U_\Phi]_{11} = F$	matches UZ_Π of Chapter 6
W (solvers)	$W(x) = e^{i \arccos(x)X}$	$\operatorname{Im}[U_\Phi^W]_{11} = F$	symmetric, $\varphi_j^W = \varphi_{d-j}^W$; $\varphi_0^O = \varphi_0^W - \frac{\pi}{2}$, else equal
C (circuit)	$\tilde{U}_j$ without explicit Z_Π	$\operatorname{Re} P$ via a signal qubit	$\varphi_j^C = \varphi_j + (-1)^{j+1} \frac{\pi}{2}$, Eq. (7.12)

Table 7.1: Three phase conventions for the same polynomial F . Only the phases at the ends and the shifts $\pm\pi/2$ differ.

A constructive proof of sufficiency is the recipe “find the completion A of (7.6), then strip layers by (7.5)”. Run in floating point it is unstable, and the working precision must grow with d [24, 35]. Practical solvers avoid it. Requiring the symmetry $\varphi_j^W = \varphi_{d-j}^W$ removes the freedom in the completion and leaves $\tilde{d} = \lceil (d + 1)/2 \rceil$ unknown *reduced phases* Φ^R , the same number as the Chebyshev coefficients $c \in \mathbb{R}^{\tilde{d}}$ in $F = \sum_k c_k T_k$ of the definite-parity target [54, 35]. Let $\mathcal{F}$ be the map from reduced phases to the coefficients of $\operatorname{Im}[U^W]_{11}$. It sends 0 to 0 and its Jacobian at 0 is $2I$, so the *fixed-point iteration*

$$\Phi^{R,(\ell+1)} = \Phi^{R,(\ell)} - \frac{1}{2}(\mathcal{F}(\Phi^{R,(\ell)}) - c), \quad \Phi^{R,(0)} = 0, \quad (7.7)$$

is Newton’s method with the Jacobian frozen at the origin. It converges geometrically to one solution when $\|c\|_1 \leq 0.861$ [18], and this condition is sufficient, not necessary. For targets that violate it, optimization-based solvers are used [20, 19]. One evaluation of $\mathcal{F}$ costs $O(d^2)$ operations by running the recursion (7.4) at the $d + 1$ Chebyshev nodes, so finding the phases takes polynomial time in d . It is done once, classically, for each target polynomial.

7.1.3 Target polynomial synthesis via convex optimization

The fixed-point iteration (7.7) takes the Chebyshev coefficient vector $c \in \mathbb{R}^{\tilde{d}}$ of an admissible polynomial $F(x)$ and finds the phase factors Φ . In applications, however, the target function $g(x)$ is often specified only on a subinterval $I \subset [0, 1]$ (for example, $I = [1/\kappa, 1]$ for matrix inversion in Section 7.3.4), whereas QSP requires F to satisfy the global admissibility constraint $|F(x)| \leq 1$ on the entire signal domain $[-1, 1]$.

Following Lin and Wiebe [35], the optimal coefficient vector $c = (c_0, \dots, c_{\tilde{d}-1}) \in \mathbb{R}^{\tilde{d}}$ can be obtained by solving a min-max convex optimization problem. Let $A_j(x) = T_{2j}(x)$ if d is even and $T_{2j+1}(x)$ if d is odd, so that $F(x) = \sum_{j=0}^{\tilde{d}-1} c_j A_j(x)$ automatically satisfies the parity constraint. To ensure that $|F(x)| \leq 1 - \eta$ strictly holds on $[0, 1]$ (with safety margin $\eta > 0$) while minimizing the uniform approximation error on I , one solves

$$\begin{aligned} \min_{c \in \mathbb{R}^{\tilde{d}}, z \geq 0} \quad & z \\ \text{subject to} \quad & |F(x) - g(x)| \leq z, \quad \forall x \in I, \\ & |F(x)| \leq 1 - \eta, \quad \forall x \in [0, 1]. \end{aligned} \tag{7.8}$$

Discretizing x over a fine grid in I and $[0, 1] \setminus I$ turns (7.8) into a standard linear program (LP). The optimal c obtained from the LP satisfies the norm constraint and can then be converted into phase factors Φ via the fixed-point iteration (7.7) or unconstrained optimization [20]. This cleanly separates the problem of polynomial approximation from the problem of phase-factor synthesis.

7.1.4 Example: a cubic

Take $F(x) = 0.2 T_1(x) + 0.4 T_3(x) = 1.6x^3 - x$. It is odd of degree 3, and $\|F\|_\infty = |F(\pm 1)| = 0.6 < 1$, while its interior extrema are ± 0.304 at $x = \mp 0.456$. Started from zero, the iteration (7.7) reaches a residual below 10^{-14} in 19 steps and returns the symmetric W -phases (0.2086, 0.1132, 0.1132, 0.2086). The O -sequence with $\text{Re } P = F$ therefore has the phases

$$\Phi = (-1.3622, 0.1132, 0.1132, 0.2086). \tag{7.9}$$

Figure 7.1 shows P and $\sqrt{1-x^2}|Q|$ for these phases. The real part is the target F . The imaginary part is the completion A of (7.6), and it is large exactly where F is small: at $x = 0$ one has $P = 0$, and at $x = \frac{1}{2}$ one has $P = -0.3 + 0.954i$, so that $|P| \approx 1$ and $Q \approx 0$. The three curves satisfy $(\text{Re } P)^2 + (\text{Im } P)^2 + (1-x^2)|Q|^2 = 1$, which is condition (iii).

7.2 Quantum singular value transformation

7.2.1 From the qubit to the matrix

Let $U \in \text{BE}_{1,m}(A)$ with singular value decomposition $A = \sum_i \sigma_i |w_i\rangle\langle v_i|$, and let Π , Z_Π and the planes $\mathcal{H}_i$, $\mathcal{H}'_i$ be as in Section 6.5.2. In the bases fixed there, U on $\mathcal{H}_i$ and $U^\dagger$ on $\mathcal{H}'_i$ are both the reflection matrix of (6.24) with λ_i replaced by σ_i , and $Z_\Pi = \text{diag}(1, -1)$ on both planes. Consequently UZ_Π (from $\mathcal{H}_i$ to $\mathcal{H}'_i$) and $U^\dagger Z_\Pi$ (from $\mathcal{H}'_i$ to $\mathcal{H}_i$) are in these bases the signal rotation $O(\sigma_i)$ of (7.1), and $e^{i\varphi_0 Z_\Pi}$ is the phase gate $e^{i\varphi Z}$. Replacing $O(x)$ in (7.2) by $\tilde{U}_j Z_\Pi$ therefore defines the *QSVT sequence*

$$U_\Phi^{\text{QSVT}} = e^{i\varphi_0 Z_\Pi} \prod_{j=1}^d \left(\tilde{U}_j Z_\Pi e^{i\varphi_j Z_\Pi} \right), \quad \tilde{U}_j = \begin{cases} U, & d-j \text{ even,} \\ U^\dagger, & d-j \text{ odd.} \end{cases} \tag{7.10}$$

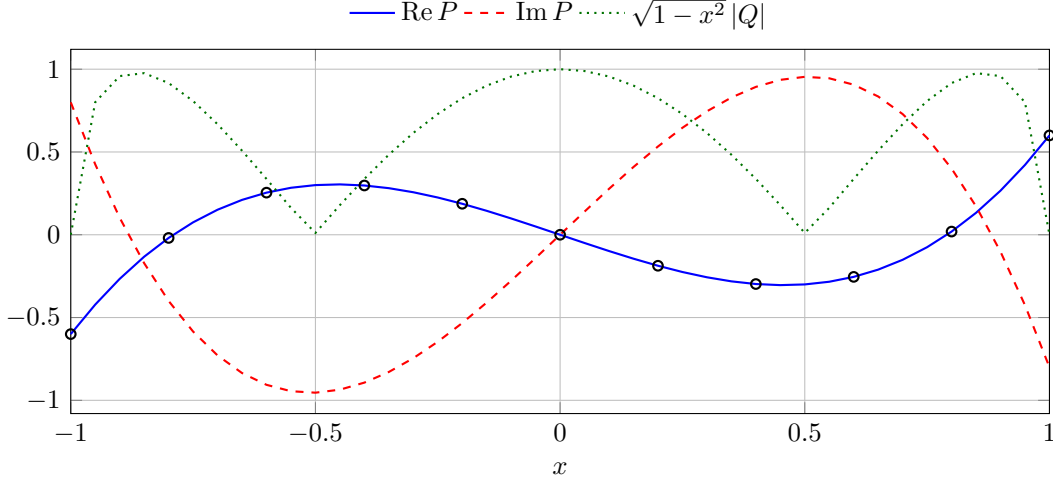


Figure 7.1: Quantum signal processing for $F(x) = 1.6x^3 - x$ with the phases (7.9). The real part of P is F (circles mark values of F); the imaginary part and $\sqrt{1-x^2}|Q|$ complete the unitary matrix (7.3), and the squares of the three curves sum to one.

The unitaries U and $U^\dagger$ alternate, and U is applied first.

Theorem (QSVT) [23, 35]. For every $\Phi \in \mathbb{R}^{d+1}$,

$$(\langle 0^m | \otimes I) U_\Phi^{\text{QSVT}} (|0^m\rangle \otimes I) = P_\Phi^{\text{SV}}(A), \quad (7.11)$$

where P_Φ is the polynomial of (7.3) and P^{SV} is the singular value transformation (6.30).

Proof. The input $|0^m\rangle|v_i\rangle$ is the first basis vector of $\mathcal{H}_i$. On the planes U_Φ^{QSVT} acts as the 2×2 matrix $U_\Phi(\sigma_i)$, and the output lies in $\mathcal{H}_i$ for even d and in $\mathcal{H}'_i$ for odd d , with first basis vector $|0^m\rangle|v_i\rangle$ or $|0^m\rangle|w_i\rangle$ respectively. The matrix element between them is the top-left entry $P_\Phi(\sigma_i)$. Summing over i gives $\sum_i P_\Phi(\sigma_i)|v_i\rangle\langle v_i|$ for even d and $\sum_i P_\Phi(\sigma_i)|w_i\rangle\langle v_i|$ for odd d , which is (6.30). $\square$

The theorem is the reason for the structure of this chapter. The same $d+1$ phases act on every plane, so they do not depend on A , on its size, or on its singular vectors, which the algorithm never needs to know. The ancilla count is that of U , and the query count is d . What remains is to make the circuit implementable and to read out a *real* polynomial.

7.2.2 The circuit

The sequence (7.10) contains Z_Π between $\tilde{U}_j$ and the phase gate, and its top-left block is the complex polynomial P_Φ . Two modifications repair both points. The first removes Z_Π by shifting the phases. Because $Z_\Pi e^{i\varphi Z_\Pi} = -i e^{i(\varphi+\pi/2)Z_\Pi} = i e^{i(\varphi-\pi/2)Z_\Pi}$, using the first form for odd j and the second for even j turns every factor into $\tilde{U}_j e^{i\varphi_j^C Z_\Pi}$ times the constant $-i$ or i , with

$$\varphi_j^C = \varphi_j + (-1)^{j+1} \frac{\pi}{2} \quad (1 \leq j \leq d), \quad \varphi_0^C = \begin{cases} \varphi_0 - \pi/2, & d \text{ odd}, \\ \varphi_0, & d \text{ even}. \end{cases} \quad (7.12)$$

The constants alternate and cancel in pairs. For odd d one factor $-i$ is left, and it is absorbed into φ_0^C because $-i e^{i\varphi_0 Z_\Pi} = Z_\Pi e^{i(\varphi_0-\pi/2)Z_\Pi}$ and $\Pi Z_\Pi = \Pi$. The operator $V_\Phi = e^{i\varphi_0^C Z_\Pi} \prod_{j=1}^d (\tilde{U}_j e^{i\varphi_j^C Z_\Pi})$ has therefore the same Π -block (7.11) as U_Φ^{QSVT} .

The second modification produces the real part. Add one qubit, the *signal qubit*, and replace every $e^{i\varphi Z_\Pi}$ by the controlled rotation

$$\text{CR}(\varphi) = \text{C}_\Pi \text{NOT} (e^{-i\varphi Z} \otimes I) \text{C}_\Pi \text{NOT} = e^{i\varphi Z \otimes Z_\Pi} = |0\rangle\langle 0| \otimes e^{i\varphi Z_\Pi} + |1\rangle\langle 1| \otimes e^{-i\varphi Z_\Pi}, \quad (7.13)$$

where $\text{C}_\Pi \text{NOT}$ flips the signal qubit when the ancilla register is in $|0^m\rangle$, an m -controlled NOT. On the signal state $|0\rangle$ the resulting circuit applies V_Φ , and on $|1\rangle$ it applies the sequence with all phases negated. In every plane the matrices of $\tilde{U}_j$ and Z_Π are real, so negating the phases conjugates the 2×2 matrix, and the top-left entry of the second branch is $\overline{P_\Phi}$. Hadamard gates before and after average the two branches, which gives the real part.

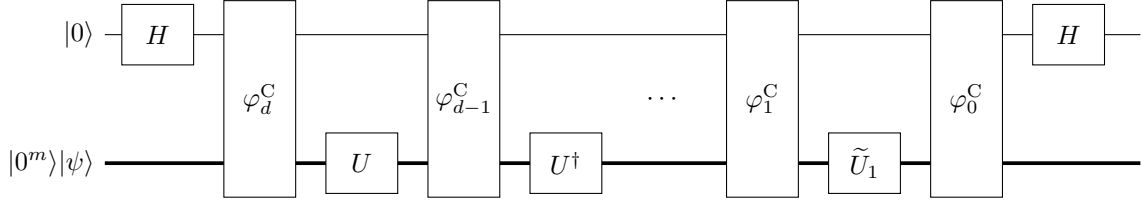


Figure 7.2: Circuit for $(\text{Re } P_\Phi)^{\text{SV}}(A)$. Time runs from left to right. The upper wire is the signal qubit; the lower wire stands for the m ancillas of U and the system. Each tall box is the controlled rotation $\text{CR}(\varphi_j^C)$ of (7.13), acting on both wires. The circuit has d uses of U or $U^\dagger$ (alternating, with $\tilde{U}_1 = U$ for odd d and $U^\dagger$ for even d) and $d+1$ rotations; the block with the signal qubit and the ancillas in $|0\rangle|0^m\rangle$ is $(\text{Re } P_\Phi)^{\text{SV}}(A)$.

Corollary (QSVT circuit) [23]. Let $U \in \text{BE}_{1,m}(A)$ and let $F \in \mathbb{R}[x]$ have degree at most d , parity $d \bmod 2$, and $|F| \leq 1$ on $[-1, 1]$. If Φ has $\text{Re } P_\Phi = F$ (Section 7.1.2), then the circuit of Figure 7.2 with the phases (7.12) is a $(1, m+1)$ -block-encoding of the singular value transformation $F^{\text{SV}}(A)$ of (6.30), with output vectors $|v_i\rangle$ for even F and $|w_i\rangle$ for odd F . It uses d queries to U and $U^\dagger$, $d+1$ gates CR , and two Hadamard gates. Each CR costs two m -controlled NOT gates and one single-qubit rotation, that is $O(m)$ elementary gates. Compared with the product rule of Section 6.2, which needs a new ancilla register for each factor, the only addition to the m ancillas of U is the signal qubit.

7.2.3 Hermitian matrices, mixed parity, and robustness

Hermitian matrices. If $A = A^\dagger$ has eigenvalues λ_i , then $|w_i\rangle = \text{sgn}(\lambda_i)|v_i\rangle$ and $\sigma_i = |\lambda_i|$. For a real F of definite parity $F(\sigma_i) \text{sgn}(\lambda_i)^d = F(\lambda_i)$, so $F^{\text{SV}}(A) = F(A)$ and QSVT becomes an eigenvalue transformation.

Mixed parity. A real polynomial p with $|p| \leq 1$ on $[-1, 1]$ splits as $p = p_e + p_o$ with $p_e(x) = \frac{1}{2}(p(x) + p(-x))$ and $p_o(x) = \frac{1}{2}(p(x) - p(-x))$, and $|p_e|, |p_o| \leq 1$. For Hermitian A the LCU rule of Section 6.3 with weights 1, 1 combines the two circuits into a $(2, m+2)$ -block-encoding of $p(A)$. If the two degrees differ by one, the shorter circuit is the beginning of the longer one, so the two share their queries, and the combined circuit needs d queries to U and $U^\dagger$ and one controlled use of U or $U^\dagger$ for the last factor of the longer circuit [35]. The factor 2 is the price of the missing parity; whether it can be reduced depends on p .

Robustness. The circuit is a product of d factors $\tilde{U}_j$ and $d+1$ phase gates, all unitary, so errors add. If each $\tilde{U}_j$ is replaced by a unitary within δ , the result moves by at most $d\delta$ in operator norm, and if each phase is off by η it moves by at most $(d+1)\eta$. If instead U encodes a matrix $\tilde{A}$ with $\|\tilde{A} - A\| \leq \epsilon'$, the perturbation theory of singular value transformations bounds $\|F^{\text{SV}}(\tilde{A}) - F^{\text{SV}}(A)\|$

by the Lipschitz constant of F times ϵ' and a polylogarithmic factor in $1/\epsilon'$ [35], and the Lipschitz constant is at most d^2 by Markov's inequality for $|F| \leq 1$.

Variants. The construction extends to projected unitary encodings $A = \tilde{\Pi}U\Pi$ with different projectors on the two sides, to rectangular A , and to controlled versions, by replacing Z_Π with $Z_{\tilde{\Pi}}$ at alternate steps [23]. Table 7.2 summarizes the resources used in this report.

target	result	queries to $U, U^\dagger$	ancillas beyond m
$F^{\text{SV}}(A)$, F real, definite parity, $\deg F = d$	$\text{BE}_{1,m+1}$	d	1
$p(A)$, $A = A^\dagger$, p real of mixed parity, $\deg p = d$	$\text{BE}_{2,m+2}$	d and one controlled U	2
A^d by the product rule (Section 6.2)	$\text{BE}_{1,dm}$	d	$(d-1)m$

Table 7.2: Resources of QSVT for a degree- d polynomial. The last row is the cost of the monomial x^d by the product rule, for comparison.

7.3 Example targets: inverse, exponential, and oscillatory filters

Sections 7.1 and 7.2 reduce “apply $f(A)$ ” for a block-encoded A with spectrum in $[-1, 1]$ to three steps: find a real polynomial p of definite parity (or two of them) with $|p| \leq 1$ on $[-1, 1]$ and $|p - f| \leq \epsilon$ on the spectrum; compute its phases; run the circuit with $d = \deg p$ queries. The cost is d times the cost of one query, times the amplification that undoes the subnormalization of the output. All the mathematics of “which function, which accuracy” sits in the first step. This section records what that step gives for the three functions named at the start of the chapter.

7.3.1 Chebyshev expansions and the Bernstein ellipse

The tool for step one is the Chebyshev series, which is a Fourier series in disguise. Put $x = \cos \theta$, so that $\theta \mapsto f(\cos \theta)$ is even and 2π -periodic, and expand

$$f(x) = a_0 + 2 \sum_{k=1}^{\infty} a_k T_k(x), \quad a_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\cos \theta) e^{-ik\theta} d\theta = \frac{1}{\pi} \int_0^{\pi} f(\cos \theta) \cos k\theta d\theta, \quad (7.14)$$

where T_k and its orthogonality (5.47) are those of Sections 3.5.2 and 5.4. Since $|T_k| \leq 1$ on $[-1, 1]$, the truncation $f_d = a_0 + 2 \sum_{k \leq d} a_k T_k$ has supremum error at most $2 \sum_{k > d} |a_k|$. How fast a_k decays is a question about analyticity, and Cauchy's integral formula answers it.

For $\rho > 1$ let $E_\rho = \{\frac{1}{2}(w + w^{-1}) : |w| = \rho\}$ be the ellipse with foci ± 1 and semi-axes $\frac{1}{2}(\rho \pm \rho^{-1})$ (the Bernstein ellipse). Suppose f is analytic in the closed region bounded by E_ρ with $|f| \leq M$ there. The function $g(w) = f(\frac{1}{2}(w + w^{-1}))$ is analytic on the annulus $\rho^{-1} \leq |w| \leq \rho$, it satisfies $g(e^{i\theta}) = f(\cos \theta)$, and its Laurent coefficients are the a_k of (7.14), with $a_{-k} = a_k$. Cauchy's estimate on the circle $|w| = \rho$ gives

$$|a_k| \leq M\rho^{-k}, \quad \|f - f_d\|_\infty \leq 2M \sum_{k > d} \rho^{-k} = \frac{2M\rho^{-d}}{\rho - 1}. \quad (7.15)$$

This is the standard estimate for Chebyshev series [52]. The size of the coefficients is thus controlled by the growth of f on ellipses around the interval, and the rate by the nearest singularity: this

is the sense in which the polynomial surrogates of Sec. 11.9 inherit their quality from the Cauchy integral. For an entire function every ρ is allowed, and the best ρ balances $M(\rho)$ against ρ^{-k} . The two entire functions below show the two typical outcomes.

7.3.2 Oscillatory filters

Hamiltonian simulation of H with $\|H\| \leq \alpha$ asks for $e^{-i\tau x}$ on $[-1, 1]$ with $\tau = \alpha t$. The function has modulus one on the real line and modulus $e^{\tau \operatorname{Im} x}$ off it, so on E_ρ with $\rho = e^s$ one has $|f| \leq e^{\tau \sinh s}$ and $|a_k| \leq \exp(\tau \sinh s - ks)$ for every $s > 0$. For $k \leq \tau$ the exponent is minimal at $s = 0$ and the bound is 1: no decay. For $k = \tau(1 + \eta)$ with small $\eta > 0$ the choice $s = \sqrt{2\eta}$ gives $|a_k| \lesssim \exp(-\frac{2\sqrt{2}}{3}\tau\eta^{3/2})$. The bound shows no decay for $k \lesssim \tau$, and beyond that the coefficients fall faster than any exponential, over a window of width of order $\tau^{1/3}$. Equating the exponent with $-\ln(1/\epsilon)$ gives the degree

$$d \approx \tau + 1.04 \tau^{1/3} \ln^{2/3}(1/\epsilon). \quad (7.16)$$

The coefficients themselves are Bessel functions. The Jacobi–Anger expansion [43] reads

$$e^{-i\tau x} = J_0(\tau) + 2 \sum_{k=1}^{\infty} (-i)^k J_k(\tau) T_k(x), \quad (7.17)$$

with J_k the Bessel functions of the first kind (Section 3.5.5). Taking real and imaginary parts, the target splits into an even and an odd polynomial, each of definite parity and bounded by one,

$$\cos(\tau x) = J_0(\tau) + 2 \sum_{k \geq 1} (-1)^k J_{2k}(\tau) T_{2k}(x), \quad \sin(\tau x) = 2 \sum_{k \geq 0} (-1)^k J_{2k+1}(\tau) T_{2k+1}(x), \quad (7.18)$$

which are the two polynomials to which QSVT is applied in Section 8.3. The exact tail $2 \sum_{k > d} |J_k(\tau)|$ gives the degrees of Table 7.3. For $\epsilon = 10^{-6}$ the excess of the degree over τ is 12, 25, 54 and 116 for $\tau = 10, 10^2, 10^3, 10^4$, which grows by about $10^{1/3}$ per decade as (7.16) predicts; that formula overestimates the excess by about ten percent.

7.3.3 Exponential filters

The heat semigroup of a positive semidefinite A with $\|A\| \leq \alpha$ is $e^{-tA} = e^{-\beta x}$ evaluated at $x = A/\alpha \in [0, 1]$, with $\beta = \alpha t$. Directly, $e^{-\beta x}$ has no parity, and the even extension $e^{-\beta|x|}$ has a kink at the origin, which polynomials resolve only slowly. The way out is to move the interval: with $y = 1 - 2x \in [-1, 1]$,

$$e^{-\beta x} = e^{-z(1-y)} = e^{-z} e^{zy}, \quad z = \frac{\beta}{2} = \frac{\alpha t}{2}, \quad (7.19)$$

which is smooth in y . It asks for a block encoding of $y = I - 2A/\alpha$ instead of A/α , and for the stencil matrices of Chapter 6 this is free. The periodic Laplacian satisfies $A^{\text{Per}} = \frac{2}{h^2}(I - H')$ with the averaging (Jacobi) operator $H' = \frac{1}{2}(S + S^\dagger)$, and H' is the two-term LCU of S and $S^\dagger$, a $(1, 1)$ -block-encoding with a single selector qubit. With $\alpha = 4/h^2$ the parameter is $z = 2t/h^2$, which is the number of explicit Euler steps t/k at the stability limit $k = h^2/2$ of (5.131) for unit diffusivity. In d dimensions H' is the average over the $2d$ neighbours, $\alpha = 4d/h^2$ and $z = 2dt/h^2$.

On E_ρ with $\rho = 1 + \eta$ one has $|e^{-z(1-w)}| = e^{z(\operatorname{Re} w - 1)} \leq \exp(z(\rho - 1)^2/(2\rho))$. The coefficients are $a_k = e^{-z} I_k(z)$ with the modified Bessel functions I_k , from the generating function $e^{(z/2)(u+1/u)} = \sum_k I_k(z) u^k$ [43]. The estimate (7.15) gives $|a_k| \leq \exp(z\eta^2/2 - k\eta)$ to leading order, which is smallest

at $\eta = k/z$ and equals $e^{-k^2/(2z)}$. The coefficients are therefore Gaussian from the start, with no plateau, and the degree is

$$d \approx \sqrt{2z \ln(1/\epsilon)}. \quad (7.20)$$

Table 7.3 lists the exact values. A dissipative semigroup can be fast-forwarded in this sense: the degree grows as $\sqrt{z}$ against the z explicit steps. The function $e^{-z(1-y)}$ has no definite parity, so Section 7.2.3 applies: $e^{-z} \cosh(zy)$ and $e^{-z} \sinh(zy)$ are run separately and combined by LCU, which gives a $(2, m+2)$ -block-encoding of e^{-tA} . Section 8.5 returns to what this does and does not provide for the heat equation.

	$e^{-i\tau x}$, parameter τ			$e^{-z(1-y)}$, parameter z		
	$\epsilon = 10^{-3}$	10^{-6}	10^{-12}	$\epsilon = 10^{-3}$	10^{-6}	10^{-12}
10	17	22	30	11	17	26
10^2	115	125	142	33	49	73
10^3	1031	1054	1090	104	155	226
10^4	10067	10116	10193	329	489	713

Table 7.3: Smallest degree d with $2 \sum_{k>d} |a_k| \leq \epsilon$, where $a_k = (-i)^k J_k(\tau)$ for the oscillatory filter and $a_k = e^{-z} I_k(z)$ for the exponential filter. The oscillatory degree is about τ plus a slowly growing excess; the exponential degree scales as $\sqrt{z}$.

7.3.4 Inverse filters

The function $1/x$ is singular at the centre of the interval, and the Bernstein argument does not apply to it directly. The solution of Childs, Kothari and Somma [14] is a polynomial that agrees with $1/x$ away from the origin and is regular at the origin. For an integer b let

$$f_b(x) = \frac{1 - (1 - x^2)^b}{x} = x \sum_{j=0}^{b-1} (1 - x^2)^j. \quad (7.21)$$

It is an odd polynomial of degree $2b - 1$. For $\kappa^{-1} \leq |x| \leq 1$ one has $|f_b(x) - 1/x| = (1 - x^2)^b/|x| \leq \kappa e^{-b/\kappa^2}$, which is at most ϵ for $b = \kappa^2 \ln(\kappa/\epsilon)$. Its Chebyshev coefficients are explicit, $f_b(x) = \sum_{j=0}^{b-1} c_{2j+1} T_{2j+1}(x)$ with

$$c_{2j+1} = 4(-1)^j \sum_{i=j+1}^b \frac{\binom{2b}{b+i}}{4^b}, \quad (7.22)$$

and they decay like a Gaussian in $j/\sqrt{b}$. Truncating at $j \leq D = \lceil \sqrt{b \ln(4b/\epsilon)} \rceil$ changes f_b by at most ϵ , so the polynomial has degree $2D + 1 = O(\kappa \log(\kappa/\epsilon))$. This count is the quantity entering the QLSP of Chapter 9.

The bound $|p| \leq 1$ needs a normalization. Near $x = c/\sqrt{b}$ one has $f_b \approx \sqrt{b}(1 - e^{-c^2})/c$, whose maximum over c is $0.638\sqrt{b}$ at $c \approx 1.12$, so $\max |f_b| \approx 0.638 \kappa \sqrt{\ln(\kappa/\epsilon)}$. The polynomial $p = f_b/(\beta \kappa)$ with $\beta \approx 0.64 \sqrt{\ln(\kappa/\epsilon)}$ is bounded by one and approximates $1/(\beta \kappa x)$. Table 7.4 shows that β is about 2 for $\kappa \leq 100$ and $\epsilon = 10^{-3}$ but grows with κ/ϵ . Polynomials with $\beta = O(1)$ and degree $O(\kappa \log(1/\epsilon))$ exist [23], and they can be found by convex optimization with fewer queries than the plain truncation. For $\kappa = 5$ and $\beta = 2$ an odd polynomial of degree 41 bounded by one on $[-1, 1]$ has $\beta \kappa p$ within 1.2×10^{-3} of $1/x$ on $[1/\kappa, 1]$, and degree 61 gives 2×10^{-5} , against 75

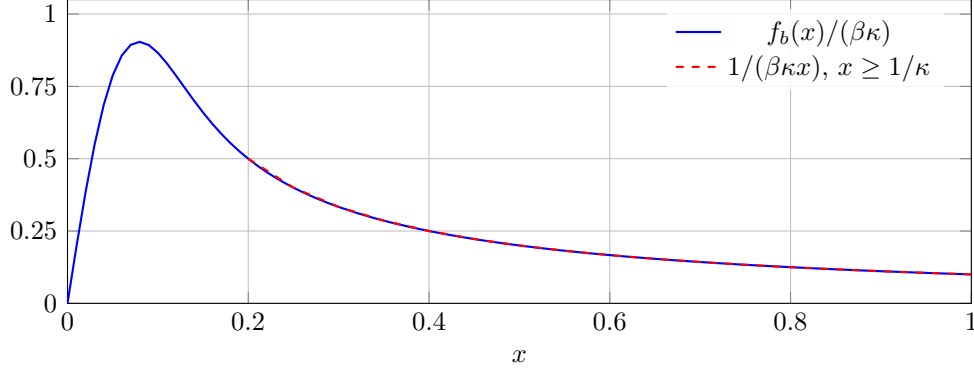


Figure 7.3: The polynomial (7.21) with $b = 200$, divided by $\beta\kappa = 10$ ($\kappa = 5$, $\beta = 2$), together with the target $1/(\beta\kappa x)$ on $[1/\kappa, 1]$. The polynomial has degree 399, is bounded by 0.904 on $[0, 1]$, and $\beta\kappa$ times it approximates $1/x$ to within 1.4×10^{-3} for $x \geq 1/\kappa$.

κ	b	degree $2D + 1$	smallest degree reaching ϵ	$\max g /\kappa$
5	213	109	75	1.87
10	922	239	163	1.94
20	3962	515	357	2.01
50	27050	1417	1061	2.10
100	115130	3033	2121	2.17

Table 7.4: The truncated polynomial g of (7.21) for $\epsilon = 10^{-3}$, with $b = \lceil \kappa^2 \ln(\kappa/\epsilon) \rceil$ and $D = \lceil \sqrt{b \ln(4b/\epsilon)} \rceil$. The last two columns give the smallest odd truncation degree with error at most ϵ on $[1/\kappa, 1]$, and the ratio $\max_{[0,1]} |g|/\kappa$, which plays the role of β .

for the truncation at $\epsilon = 10^{-3}$. Figure 7.3 shows the shape: the polynomial follows $1/(\beta\kappa x)$ for $|x| \geq 1/\kappa$, bends over below it, and returns to zero at the origin as an odd function must.

The linear dependence on κ cannot be avoided by a better choice of polynomial. Suppose p is bounded by one on $[-1, 1]$ and within ϵ' of $1/(\beta\kappa x)$ on $[1/\kappa, 1]$. Between $x = 1/\kappa$ and $x = 2/\kappa$ the target drops from $1/\beta$ to $1/(2\beta)$, so by the mean value theorem $|p'(\xi)| \geq \kappa(\frac{1}{2\beta} - 2\epsilon')$ for some $\xi \in (1/\kappa, 2/\kappa)$. Bernstein's inequality $|p'(\xi)| \leq d/\sqrt{1 - \xi^2}$ for $|p| \leq 1$ gives $d \geq \frac{\sqrt{3}}{2}\kappa(\frac{1}{2\beta} - 2\epsilon') = \Omega(\kappa/\beta)$ for $\kappa \geq 4$ and $\epsilon' < 1/(4\beta)$. The sign function is the other standard target. It is approximated on $[-1, -\delta] \cup [\delta, 1]$ by an odd polynomial of degree $O(\delta^{-1} \log(1/\epsilon))$ built from the error function [35], and it gives projectors onto spectral windows. It is not needed in this report.

7.3.5 Summary

Table 7.5 lists the three targets with the structure that decides their cost: the part of $[-1, 1]$ on which the approximation is needed, the parity, the degree, and the chapter that uses them. The oscillatory filter costs a degree linear in $\tau = \alpha t$ plus a lower-order term, and the linear dependence cannot be improved (Chapter 8, no-fast-forwarding theorem); the exponential filter costs $\sqrt{\tau \log(1/\epsilon)}$ because the semigroup damps high frequencies; the inverse filter costs κ because it must resolve the smallest singular value.

target	needed on	parity	degree	used in
$\cos(\tau x), \sin(\tau x)$	$[-1, 1]$	even, odd	$\tau + O(\tau^{1/3} \log^{2/3}(1/\epsilon))$	Chapter 8
$e^{-z(1-y)}$	$[-1, 1]$	none (split)	$O(\sqrt{z \log(1/\epsilon)})$	Sec. 8.5, Chapter 14
$1/(\beta \kappa x)$	$[1/\kappa, 1]$	odd	$O(\kappa \log(\kappa/\epsilon))$	Chapter 9

Table 7.5: The three polynomial targets of this chapter.

Chapter 8

Hamiltonian simulation

Hamiltonian simulation implements e^{-iHt} (or a related semigroup) given a block encoding or a sparse description of H . In this report it sits *after* block encoding and QSP/QSVT, because those are the primitives from which modern simulation algorithms are built [8, 40, 35].

Simulating quantum dynamics was the original motivation for quantum computers [22]. Throughout, $H = H^\dagger$ acts on n qubits, and the input is either a description of H as a short sum of terms whose exponentials are easy, or a block encoding $U_H \in \text{BE}_{\alpha,m}(H)$ in the sense of Chapter 6. Given a time t and an accuracy ϵ , the task is a circuit W with $\|W - e^{-iHt}\| \leq \epsilon$, or a block encoding of e^{-iHt} of that accuracy. The cost is counted in queries to U_H and $U_H^\dagger$, or in exponentials of the terms, together with the extra gates and ancilla qubits. Since $e^{-iHt} = e^{-i(H/\alpha)\alpha t}$, the normalization can be absorbed into the time, and the cost depends on H only through the dimensionless product $\tau = \alpha t$.

Two lower bounds frame what follows. The first is the no-fast-forwarding theorem: for a generic sparse H any algorithm needs $\Omega(\tau)$ queries [8]. Unitary evolution cannot be compressed in time, in contrast with the heat semigroup of Section 7.3.3. The second concerns precision: $\Omega(\log(1/\epsilon)/\log \log(1/\epsilon))$ queries are needed even for short times [10]. An algorithm is *optimal* if its cost is $O(\tau + \log(1/\epsilon)/\log \log(1/\epsilon))$, and Section 8.3 shows that this can be attained.

The generators that appear in this report are collected in Table 8.1. They come from the periodic stencils of Chapter 5 and are written with the cyclic shift S , the central difference $D_0 = (S^\dagger - S)/(2h)$, and the Laplacian $A^{\text{Per}} = (2I - S - S^\dagger)/h^2$ of Section 5.9.1. The first two rows are Hermitian and are the subject of Sections 8.1–8.3; the last row is not, and Section 8.5 turns to it.

equation (periodic, 1D)	generator	scale α	treated in
Schrödinger $iu_t = -u_{xx}$	$H = A^{\text{Per}} \succeq 0, u_t = -iHu$	$4/h^2$	Secs. 8.1, 8.3
transport $u_t + cu_x = 0$	$H = -icD_0 = \frac{ic}{2h}(S - S^\dagger)$	c/h	Sec. 8.3, Ch. 13
wave $u_{tt} = c^2u_{xx}$	first-order system, Hermitian after the factorization of Ch. 13	$O(c/h)$	Ch. 13
heat $u_t = u_{xx}$	$A = A^{\text{Per}}$, semigroup e^{-tA} (not unitary)	$4/h^2$	Sec. 8.5, Ch. 14

Table 8.1: Generators of the PDEs treated in this report, in the periodic one-dimensional setting. The normalization scale α is that of the shift-based block encoding of Section 6.4.1; $\tau = \alpha t$ is the cost parameter.

8.1 Trotter–Suzuki product formulae

Suppose $H = H_1 + H_2$, where $e^{-iH_1\tau}$ and $e^{-iH_2\tau}$ can be implemented but $e^{-iH\tau}$ cannot, because $[H_1, H_2] \neq 0$. The Lie–Trotter formula [53], $e^{-i(H_1+H_2)t} = \lim_{r \rightarrow \infty} (e^{-iH_1 t/r} e^{-iH_2 t/r})^r$, alternates the two exponentials over short steps, and it gave the first quantum algorithm for local Hamiltonians [38]. It needs no block encoding, no ancilla qubit and no amplitude amplification, only the ability to exponentiate the pieces. The price is paid in the number of steps.

8.1.1 First order

Let $S_1(\tau) = e^{-iH_1\tau} e^{-iH_2\tau}$ and $H = H_1 + H_2$. Then

$$\|S_1(\tau) - e^{-iH\tau}\| \leq \frac{\tau^2}{2} \|[H_1, H_2]\|. \quad (8.1)$$

Proof. Put $F(s) = e^{-iH_1 s} e^{-iH_2 s}$ and $G(s) = e^{-iH s}$. Differentiating gives $F'(s) = -iH F(s) + E(s)F(s)$ with $E(s) = -i(e^{-iH_1 s} H_2 e^{iH_1 s} - H_2)$. Since $\frac{d}{ds}[G(\tau - s)F(s)] = G(\tau - s)E(s)F(s)$, Duhamel’s formula (variation of constants) reads $F(\tau) - G(\tau) = \int_0^\tau G(\tau - s)E(s)F(s) ds$. Because $\frac{d}{dr}(e^{-iH_1 r} H_2 e^{iH_1 r}) = -i e^{-iH_1 r} [H_1, H_2] e^{iH_1 r}$, one has $\|E(s)\| \leq s \|[H_1, H_2]\|$, and integrating over $s \in [0, \tau]$ gives (8.1). $\square$

Over r steps of length t/r the errors of the unitary factors add, so

$$\|S_1(t/r)^r - e^{-iHt}\| \leq \frac{t^2}{2r} \|[H_1, H_2]\|, \quad r = \left\lceil \frac{t^2 \|[H_1, H_2]\|}{2\epsilon} \right\rceil. \quad (8.2)$$

The number of steps is quadratic in t and linear in $1/\epsilon$, and it is proportional to the commutator: if H_1 and H_2 commute the formula is exact. With $\alpha = \|H_1\| + \|H_2\|$ the worst case is $\|[H_1, H_2]\| \leq 2\|H_1\|\|H_2\| \leq \alpha^2/2$, and $r = O(\tau^2/\epsilon)$.

8.1.2 Higher order

Symmetrizing removes the first-order term. The Strang formula $S_2(\tau) = e^{-iH_1\tau/2} e^{-iH_2\tau} e^{-iH_1\tau/2}$ satisfies [15]

$$\|S_2(\tau) - e^{-iH\tau}\| \leq \frac{\tau^3}{12} \|[H_2, [H_2, H_1]]\| + \frac{\tau^3}{24} \|[H_1, [H_1, H_2]]\|, \quad (8.3)$$

so $r = O(t^{3/2}\epsilon^{-1/2})$ steps suffice. Suzuki [48, 49] builds formulas of every even order recursively,

$$S_{2k}(\tau) = S_{2k-2}(u_k\tau)^2 S_{2k-2}((1 - 4u_k)\tau) S_{2k-2}(u_k\tau)^2, \quad u_k = \frac{1}{4 - 4^{1/(2k-1)}}, \quad (8.4)$$

which has order $2k$ and uses 5^{k-1} Strang stages. The middle step $(1 - 4u_k)\tau$ is negative, so that stage runs backward in time. Yoshida’s triple jump [55], $S_4(\tau) = S_2(s\tau)S_2((1 - 2s)\tau)S_2(s\tau)$ with $s = 1/(2 - 2^{1/3})$, is a three-stage fourth-order variant.

A formula of order p has error $O(\tau^{p+1})$ per step, hence $r = O(t^{1+1/p}\epsilon^{-1/p})$ steps. For $H = \sum_{\gamma=1}^\Gamma H_\gamma$ the commutator-scaling theorem [15, 35] makes this precise: a p -th order formula with K stages per step reaches accuracy ϵ with

$$r = O\left((tK)^{1+1/p} \|H\|_{p+1, \text{comm}}^{1/p} \epsilon^{-1/p}\right) \quad (8.5)$$

steps, where $\|H\|_{p+1, \text{comm}} = \sum_{\gamma_1, \dots, \gamma_{p+1}} \|[H_{\gamma_{p+1}}, [\dots, [H_{\gamma_2}, H_{\gamma_1}]\dots]]\|$ sums the norms of the nested commutators of order $p+1$, and $K = 2 \cdot 5^{p/2-1}$ for Suzuki’s formulas. The exponent $1 + 1/p$ tends to 1 for large p , but the constants grow like $5^{p/2}$ and no fixed order reaches the linear dependence on t that the lower bound permits.

8.1.3 Example: the periodic Laplacian split into even and odd bonds

Take $N = 2^n$ sites, so that the site index is a binary number whose lowest bit is the parity of the site. The periodic Laplacian is the sum over the N bonds $(j, j+1)$ of the one-bond operators $h^{-2}(|j\rangle - |j+1\rangle)(\langle j| - \langle j+1|)$. Grouping the bonds with even and odd j (Figure 8.1) gives

$$A^{\text{Per}} = A_e + A_o, \quad A_e = \frac{I - X_0}{h^2}, \quad A_o = \frac{I - SX_0S^\dagger}{h^2}, \quad (8.6)$$

where X_0 is the Pauli X on the lowest qubit, which swaps the sites $2m$ and $2m+1$, and the conjugation by the cyclic shift S moves the pairs by one site. Within one family the bonds do not overlap, so the terms commute and no splitting error arises inside A_e or inside A_o . Since X_0 is an involution,

$$e^{-itA_e} = e^{-it/h^2} \left(\cos \frac{t}{h^2} I + i \sin \frac{t}{h^2} X_0 \right), \quad (8.7)$$

which is a single-qubit rotation, and $e^{-itA_o} = S e^{-itA_e} S^\dagger$ costs the same plus a cyclic shift and its inverse. One Trotter layer therefore costs $\text{poly}(n)$ gates, with no ancilla.

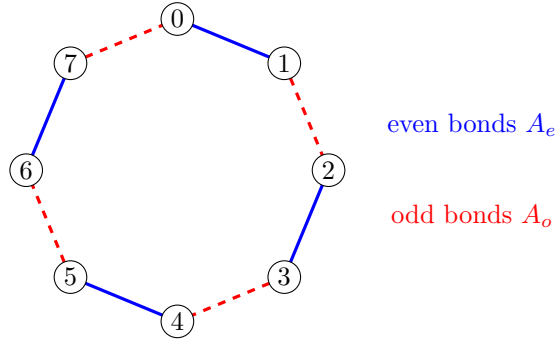


Figure 8.1: A ring of $N = 8$ sites. The periodic Laplacian is the sum of the even bonds (solid) and the odd bonds (dashed). Bonds within a family have disjoint supports and commute; the two families share sites and do not.

The error is controlled by the commutator. Since X_0 and $X_1 = SX_0S^\dagger$ are involutions, $[A_e, A_o] = h^{-4}[X_0, X_1]$, and X_0X_1 moves an even site by -2 and an odd site by $+2$, so that

$$[A_e, A_o] = \frac{1}{h^4} (S^{-2} - S^2) D, \quad D = \text{diag}((-1)^j). \quad (8.8)$$

The circulant $S^{-2} - S^2$ has eigenvalues of modulus $2|\sin(4\pi k/N)|$ and D is unitary, so $\|[A_e, A_o]\| = 2/h^4$. The same computation gives $\|A_e\| = \|A_o\| = 2/h^2$ and $\|[A_e, [A_e, A_o]]\| = \|[A_o, [A_o, A_e]]\| = 4/h^6$. With $\tau = \alpha t = 4t/h^2$, (8.2) requires $r = \lceil \tau^2/(16\epsilon) \rceil$ Lie-Trotter steps, and (8.3) gives the error $\tau^3/(128r^2)$ for r Strang steps, that is $r = \lceil \sqrt{\tau^3/(128\epsilon)} \rceil$. Each step contains two exponentials.

These are worst-case counts, valid for every initial state. For a smooth state the commutator is much smaller. By (8.8), $[A_e, A_o]u$ is $h^{-4}(-1)^j(u_{j+2} - u_{j-2}) \approx 4h^{-3}(-1)^j u'(x_j)$, so its norm is about $4\|u'\|/h^3$, smaller than $2\|u\|/h^4$ by a factor of order kh for a state of wavenumber k . The exact flow preserves such smoothness, because A^{Per} commutes with $e^{-itA^{\text{Per}}}$, and the vector-norm analysis [4, 35] shows that the improvement persists along the evolution. This gain depends on the state, not on the algorithm, and it is the main reason product formulae remain competitive for PDE problems with smooth data.

8.1.4 Exact simulation of 1-sparse Hamiltonians

For general Hamiltonians specified through sparse black-box oracles rather than analytical stencils, Trotter splitting serves as an indispensable bridge [8]. Suppose H is a d -sparse Hermitian matrix acting on n qubits, accessible via two standard oracles:

- a column oracle $O_f: |x\rangle|i\rangle \mapsto |x\rangle|f(x, i)\rangle$, where $f(x, i)$ is the column index of the i th nonzero entry in row x ($1 \leq i \leq d$);
- a matrix-entry oracle $O_H: |x\rangle|y\rangle|0\rangle \mapsto |x\rangle|y\rangle|H_{xy}\rangle$.

By an efficient edge coloring of the graph induced by H , the Hamiltonian decomposes into a sum of $m = O(d^2)$ (or $O(d)$ with sophisticated edge-coloring) strictly 1-sparse Hermitian operators, $H = \sum_{j=1}^m H_j$, where every row and column of each H_j contains at most one nonzero entry [8].

The fundamental building block is therefore the exact simulation of an individual 1-sparse block H_j over time t . Because each row of H_j has at most one nonzero entry, the dynamics in the computational basis uniquely and exhaustively partitions into disjoint invariant subspaces of dimension at most two [8]:

- *Case I (off-diagonal, $x \neq f(x)$):* The state $|x\rangle$ transitions into $|f(x)\rangle$. Within the 2D invariant subspace spanned by $\{|x\rangle, |f(x)\rangle\}$, the submatrix is

$$\begin{pmatrix} 0 & H_{x,f(x)} \\ H_{f(x),x} & 0 \end{pmatrix} = |H_{x,f(x)}| \begin{pmatrix} 0 & e^{-i\theta} \\ e^{i\theta} & 0 \end{pmatrix}, \quad (8.9)$$

where $H_{f(x),x} = H_{x,f(x)}^*$ by Hermiticity. The unitary time evolution acts as a single-qubit rotation $R_x(2|H_{x,f(x)}|t)$ accompanied by phases $e^{\pm i\theta}$.

- *Case II (diagonal, $x = f(x)$):* The state maps strictly to itself, forming an isolated 1D invariant subspace. The evolution simply applies the diagonal phase accumulation $|x\rangle \mapsto e^{-iH_{xx}t}|x\rangle$.

Crucially, these two cases can be handled coherently across the entire superposition without any classical branching or measurement [8]. By computing an index comparison $c \leftarrow (x > f(x))$ into a flag qubit and applying a Controlled-SWAP between the system and index registers, the registers are sorted to $(\min(x, f(x)), \max(x, f(x)))$. This guarantees that the matrix element $H_{\min, \max}$ is queried consistently, enforcing Hermiticity automatically. After applying the subspace rotation or diagonal phase conditioned on the parity $x \oplus f(x)$, uncomputing the registers restores all ancillas to zero. Exactly four oracle queries ($2 \times O_f$ and $2 \times O_H$) implement $e^{-iH_j t}$ exactly, and Trotter–Suzuki formulas then assemble the full simulation e^{-iHt} at order p with gate overhead governed by (8.5).

8.1.5 Strengths and limits

Product formulae are the right tool when the pieces are cheap and the accuracy is moderate: no block encoding, no ancilla, and a gate count proportional to the number of terms times the number of steps. Their weaknesses are the polynomial dependence $\epsilon^{-1/p}$ on the accuracy, which is a real handicap when the simulation is a subroutine of a linear solver and ϵ must be small (Chapter 9), the superlinear dependence $t^{1+1/p}$ on time, and, for stencil matrices, commutators of size h^{-4} at worst. The following sections remove the first two weaknesses.

8.2 Linear combination of unitaries and truncated Taylor series

Berry, Childs, Cleve, Kothari and Somma replace the splitting by a Taylor expansion realized as a linear combination of unitaries [9]. Let $H = \sum_l \alpha_l U_l$ be an LCU with unitary U_l and $\alpha = \sum_l |\alpha_l|$ as in Section 6.3, or equivalently let a block encoding U_H be given. The idea is that the Taylor series of e^{-iHt} is a combination of products of unitaries with explicit weights, and that for a short enough time its total weight is less than 2.

8.2.1 Segments and the truncated series

Split t into r segments of length t/r and approximate each by the truncated series

$$e^{-iHt/r} \approx \sum_{k=0}^K \frac{(-iHt/r)^k}{k!} = \sum_{k=0}^K \frac{(-it/r)^k}{k!} \sum_{l_1, \dots, l_k} \alpha_{l_1} \cdots \alpha_{l_k} U_{l_1} \cdots U_{l_k}. \quad (8.10)$$

The right-hand side is an LCU of products of unitaries with total weight $s_K = \sum_{k \leq K} (\alpha t/r)^k / k!$. Take $r = \lceil \alpha t / \ln 2 \rceil$, so that $\alpha t/r \leq \ln 2$ and $s_K < e^{\ln 2} = 2$. Raising the normalization α slightly, at the cost of one more ancilla qubit, makes the segment time exactly $\ln 2$ and $s_K \rightarrow 2$ as $K \rightarrow \infty$. The tail beyond order K is at most $\sum_{k > K} (\ln 2)^k / k! \leq 2(\ln 2)^{K+1} / (K+1)!$, and requiring it to be at most ϵ/r gives

$$K = O\left(\frac{\log(r/\epsilon)}{\log \log(r/\epsilon)}\right). \quad (8.11)$$

The LCU circuit PREP–SELECT–PREP[†] of Section 6.3 (with a selector register that holds k and the indices $l_1, \dots, l_k$, compressed to $O(m + \log K)$ qubits [35]) block-encodes $s_K^{-1} e^{-iHt/r}$ up to the truncation error. Since $s_K \approx 2$, it is a block encoding W of $\frac{1}{2}V$ with $V \approx e^{-iHt/r}$ unitary.

8.2.2 Oblivious amplitude amplification

Postselecting on the ancilla succeeds with probability only about 1/4 per segment, and the probabilities multiply over the r segments. Amplitude amplification cannot be applied directly because the input state is unknown, but here it is not needed: since the target block $\frac{1}{2}V$ is a *unitary* divided by two, all singular values of the block equal $\frac{1}{2}$, and three uses of W turn them into one. By the qubitization identity (6.29) with $k = 1$, the operator

$$\widetilde{W} = W Z_{\Pi} W^{\dagger} Z_{\Pi} W \quad (8.12)$$

has the block $T_3^{\text{SV}}(\frac{1}{2}V) = -V$, because $T_3(\frac{1}{2}) = 4 \cdot \frac{1}{8} - \frac{3}{2} = -1$. The final Z_{Π} in (6.29) is absent here, which does not change the Π -block since Z_{Π} acts as the identity on $|0^m\rangle$. The sign is a global phase. This is *oblivious* amplitude amplification [9]: it uses only that V is unitary and works for every input state at once. If V is only approximately unitary, the error stays proportional to the truncation error (robust oblivious amplitude amplification [10, 35]).

8.2.3 Cost

Each segment uses W three times, and each W contains products of up to K unitaries, that is K queries to U_H . The total is $3rK$ queries. With $r = O(\tau)$ and (8.11),

$$O\left(\tau \frac{\log(\tau/\epsilon)}{\log \log(\tau/\epsilon)}\right) \quad (8.13)$$

queries, with $O(m + \log \log(\tau/\epsilon))$ ancilla qubits [9, 35]. This is linear in τ up to a logarithm and logarithmic in $1/\epsilon$, a large improvement over product formulae, but the two costs are multiplied, while the lower bound of the introduction adds them. Section 8.3 closes that gap.

8.3 Qubitization and QSP-based simulation

The optimal algorithm uses the machinery of Chapter 7 with the oscillatory target of Section 7.3.2 [39, 40]. Chapter 6 showed that a block encoding of H/α is turned by qubitization into a walk whose eigenphases are $\arccos(\lambda_i/\alpha)$. QSVT turns that walk into any polynomial of H/α , and the polynomials needed for $e^{-i\tau x}$ are the Jacobi–Anger truncations.

8.3.1 The two polynomials

By (7.18), the even and odd parts of $e^{-i\tau x} = \cos(\tau x) - i \sin(\tau x)$ are Chebyshev series with Bessel coefficients. Truncating them at degree d gives

$$C_d(x) = J_0(\tau) + 2 \sum_{k=1}^{\lfloor d/2 \rfloor} (-1)^k J_{2k}(\tau) T_{2k}(x), \quad S_d(x) = 2 \sum_{k=0}^{\lfloor (d-1)/2 \rfloor} (-1)^k J_{2k+1}(\tau) T_{2k+1}(x), \quad (8.14)$$

and since $|T_k| \leq 1$ on $[-1, 1]$, each of them differs from $\cos(\tau x)$ and $\sin(\tau x)$ by at most $2 \sum_{k>d} |J_k(\tau)|$. The bound $|J_k(\tau)| \leq (\tau/2)^k/k!$ [43] already gives $d = O(\tau + \log(1/\epsilon))$, and sharper estimates of the Bessel tail give the optimal $O(\tau + \log(1/\epsilon)/\log(e + \log(1/\epsilon)/\tau))$ [35]. The exact tail decays in the way shown in Figure 8.2: it stays at order one up to $d \approx \tau$ and then falls super-exponentially, with a transition that sharpens as τ grows, as in Section 7.3.2. Table 7.3 lists the degrees.

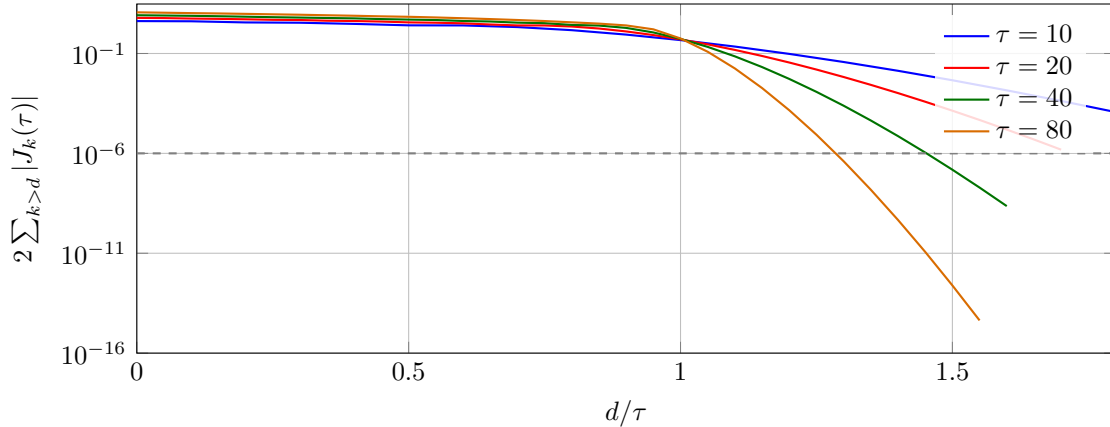


Figure 8.2: The truncation bound $2 \sum_{k>d} |J_k(\tau)|$ for the Jacobi–Anger series (7.17), against d/τ , for $\tau = 10, 20, 40, 80$. The dashed line marks $\epsilon = 10^{-6}$. The error stays at order one up to $d \approx \tau$ and then falls faster than any exponential, and the transition sharpens as τ grows.

8.3.2 The algorithm and its cost

Let $U_H \in \text{BE}_{\alpha,m}(H)$ with $H = H^\dagger$ and put $\tau = \alpha t$. The algorithm has five steps.

1. Choose the smallest d for which C_d and S_d approximate $\cos(\tau x)$ and $\sin(\tau x)$ to accuracy $\epsilon/2$ on $[-1, 1]$. Both truncations are bounded by $1 + \epsilon/2$ on $[-1, 1]$; divide them by $\beta = 1 + \epsilon$ to enforce $|p| \leq 1$.

2. Compute the phases of the even polynomial C_d/β and of the odd polynomial S_d/β by the methods of Section 7.1.2.
3. Build the QSVT circuits of Figure 7.2 for the two polynomials. They are $(1, m+1)$ -block-encodings of $C_d(H/\alpha)/\beta$ and $S_d(H/\alpha)/\beta$, and for Hermitian H the singular value transformation is the eigenvalue transformation (Section 7.2.3).
4. Combine them by an LCU with one more qubit and weights $1, -i$ (Section 7.2.3), giving a block encoding of $(C_d(H/\alpha) - iS_d(H/\alpha))/(2\beta)$.
5. Optionally amplify with (8.12), three uses of the circuit, to turn the normalization 2β into 1.

Theorem (optimal Hamiltonian simulation) [39, 40, 35]. Let $U_H \in \text{BE}_{\alpha, m}(H)$ with $H = H^\dagger$, $t > 0$ and $\epsilon \in (0, 1)$. There is a $(2(1 + \epsilon), m + 2, \epsilon)$ -block-encoding of e^{-iHt} that uses

$$O\left(\alpha t + \frac{\log(1/\epsilon)}{\log(e + \log(1/\epsilon)/(\alpha t))}\right) \quad (8.15)$$

queries to U_H and $U_H^\dagger$ and one query to the controlled U_H . After one round of oblivious amplitude amplification it is a $(1, m + 2, 9\epsilon)$ -block-encoding at three times the cost.

The cost (8.15) matches both lower bounds of the introduction, so the algorithm is optimal. Only one extra qubit is needed for the signal and one for the LCU. The normalization of the polynomials is the essential point. The expansion (8.14) could also be implemented as a linear combination of the Chebyshev blocks of Section 6.5.5, but its coefficient 1-norm $|J_0(\tau)| + 2\sum_{k \geq 1} |J_k(\tau)|$ grows like $\sqrt{\tau}$ (it is 11.6 at $\tau = 80$ and 44 at $\tau = 1280$), so amplification would add a factor of order $\sqrt{\tau}$ to the query count. The phases of QSVT absorb the whole polynomial into a normalization of one.

8.3.3 PDE test beds

Free Schrödinger equation. Take $H = A^{\text{Per}}$ with the shift-based block encoding U_L of Section 6.4.1, with $m = 2$ and $\alpha = 4/h^2$. The evolution is the free Schrödinger equation $iu_t = -u_{xx}$ on a ring. The encoded matrix is Hermitian although U_L is not, and Section 7.2.3 applies. The query count is the degree of Table 7.3 at $\tau = 4t/h^2$: 17 to 22 for $\tau = 10$, 115 to 125 for $\tau = 100$, and 1031 to 1054 for $\tau = 1000$, for ϵ between 10^{-3} and 10^{-6} .

Transport. The generator $H = -icD_0$ (with central difference $D_0 = \frac{S^\dagger - S}{2h}$) is Hermitian, with eigenvalues $(c/h) \sin(2\pi k/N)$, and it is a two-term LCU of the unitaries iS and $-iS^\dagger$ with $\alpha = c/h$ and a single selector qubit. Here $\tau = ct/h$ is the number of grid cells crossed by the wave in time t , and the algorithm needs about τ queries. This is the number of steps of an explicit scheme at Courant number one, but the state lives on $\log_2 N$ qubits and the accuracy enters only through $\log(1/\epsilon)$. The cost grows in proportion to c/h , so a finer mesh is paid for linearly in $1/h$.

Fast-forwardable cases. Constant-coefficient problems on a torus are diagonalized by the quantum Fourier transform, $e^{-itH} = F^\dagger \text{diag}(e^{-it\lambda_k})F$ with $\lambda_k = (4/h^2) \sin^2(\pi k/N)$ for the Laplacian, as in Section 5.9.2. The phases are computed from k by quantum arithmetic with a gate count independent of t . This does not contradict the lower bound, which is for generic sparse H : the Fourier transform exploits translation invariance. The algorithms of this chapter are needed when the coefficients vary, the geometry is not a torus, or only a block encoding is available.

8.3.4 Comparison

Table 8.2 summarizes the asymptotic costs and Table 8.3 gives counts for the Laplacian test bed.

method	queries or exponentials	ancilla qubits	input
Lie–Trotter	$O(\tau^2/\epsilon)$, worst case	none	exponentials of the terms
Suzuki, order p	$O(\tau^{1+1/p}\epsilon^{-1/p})$, constants $5^{O(p)}$	none	exponentials of the terms
LCU + Taylor + OAA	$O(\tau \log(\tau/\epsilon)/\log \log(\tau/\epsilon))$	$O(m + \log \log(\tau/\epsilon))$	block encoding or LCU
QSP (qubitization)	$O(\tau + \log(1/\epsilon)/\log \log(1/\epsilon))$	$m + 2$	block encoding
lower bound	$\Omega(\tau + \log(1/\epsilon)/\log \log(1/\epsilon))$		

Table 8.2: Cost of Hamiltonian simulation to accuracy ϵ , in the dimensionless time $\tau = \alpha t$.

τ	ϵ	Lie	Strang	Taylor + OAA	QSP	QSP + OAA
10	10^{-3}	1.25×10^4	178	270	17	51
10	10^{-6}	1.25×10^7	5592	405	22	66
10^2	10^{-3}	1.25×10^6	5592	3045	115	345
10^2	10^{-6}	1.25×10^9	1.77×10^5	4350	125	375
10^3	10^{-3}	1.25×10^8	1.77×10^5	34632	1031	3093
10^3	10^{-6}	1.25×10^{11}	5.59×10^6	43290	1054	3162

Table 8.3: Counts for the periodic Laplacian of Section 8.1.3. *Lie* and *Strang*: number of exponentials $e^{-i\tau'A_e}$, $e^{-i\tau'A_o}$ from the worst-case bounds (8.2) and (8.3), namely $2\lceil\tau^2/(16\epsilon)\rceil$ and $2L + 1$ with $L = \lceil\sqrt{\tau^3/(128\epsilon)}\rceil$. *Taylor + OAA*: $3rK$ queries with $r = \lceil\tau/\ln 2\rceil$ and K the smallest order with $\sum_{k>K}(\ln 2)^k/k! \leq \epsilon/r$. *QSP*: the degree of Table 7.3; *QSP + OAA*: three times that. A Trotter layer and a query to U_L both cost $\text{poly}(n)$ gates, so the columns are comparable up to constants.

8.4 Time-dependent simulation and the interaction picture

When the Hamiltonian depends explicitly on time, $H = H(t)$, an essential complication arises: operators at different times generally do not commute, $[H(t_1), H(t_2)] \neq 0$. Consequently, the naive exponent $e^{-i\int_0^t H(s) ds}$ fails to solve the Schrödinger equation $i\partial_t|\psi(t)\rangle = H(t)|\psi(t)\rangle$. This section discusses the time-ordering formalism, its implementation via truncated Dyson series, and how the interaction picture fundamentally reduces the query cost for PDE problems with stiff kinetic operators [41, 35].

8.4.1 Time-ordering and the truncated Dyson series

The exact unitary evolution operator $U(t)$ satisfying $i\partial_t U(t) = H(t)U(t)$ with $U(0) = I$ is given by the time-ordered exponential [35]

$$U(t) = \mathcal{T} \exp\left(-i \int_0^t H(s) ds\right) = I + \sum_{n=1}^{\infty} (-i)^n \int_0^t ds_n \int_0^{s_n} ds_{n-1} \cdots \int_0^{s_2} ds_1 H(s_n) H(s_{n-1}) \cdots H(s_1), \quad (8.16)$$

which is the Dyson series. The time-ordering operator $\mathcal{T}$ arranges any product of time-dependent operators in chronological order from right to left: $\mathcal{T}\{H(t_1)H(t_2)\} = H(t_1)H(t_2)$ if $t_1 > t_2$ and $H(t_2)H(t_1)$ if $t_2 > t_1$.

Low and Wiebe [41] generalized the truncated Taylor series (Section 8.2) to simulate (8.16) by LCU. Suppose $H(t)$ is differentiable and accessible via a time-dependent block encoding with

subnormalization $\alpha_T = \sup_{s \in [0, t]} \|H(s)\|$. The total duration t is divided into $r = \lceil \alpha_T t / \ln 2 \rceil$ short segments of length $\Delta t \leq \ln 2 / \alpha_T$. On each segment, the Dyson series is truncated at order

$$K = O\left(\frac{\log(r/\epsilon)}{\log \log(r/\epsilon)}\right). \quad (8.17)$$

The continuous nested integrals over $s_1 < \dots < s_n$ are discretized on a fine time grid with $n_t = O(\log(\max_s \|H'(s)\| t^2 / \epsilon))$ qubits. A coherent comparison and sorting network arranges the sampled time coordinates chronologically, allowing the time-ordered product to be implemented via a multiplexed SELECT oracle. On each segment, the resulting LCU unitary encodes $s_K^{-1} U(\Delta t)$ with $s_K \leq 2$, which is deterministically boosted to a $(1, \cdot)$ -block-encoding by a single round of oblivious amplitude amplification (Section 8.2.2).

Theorem (truncated Dyson simulation) [41, 35]. Let $H(t)$ be a differentiable time-dependent Hamiltonian on $[0, t]$ with $\alpha_T(s) = \|H(s)\|$. Truncated Dyson simulation achieves precision ϵ using

$$\tilde{O}\left(\int_0^t \alpha_T(s) ds \cdot \log(1/\epsilon)\right) \quad (8.18)$$

queries to the block encoding of $H(t)$.

Remark (The query complexity gap). Unlike time-independent simulation, where qubitization achieves the strictly additive scaling $O(\alpha t + \log(1/\epsilon) / \log \log(1/\epsilon))$ of Theorem 8.3.2, time-dependent simulation currently incurs a multiplicative scaling $\tilde{O}(\alpha_T t \log(1/\epsilon))$. This gap stems from a structural distinction: qubitization relies on the stationary eigenvalues of H to construct a single 2D invariant walk operator $W = R_0 U^\dagger R_0 U$. Time-dependent Hamiltonians have instantaneous eigenvalues but generally no stationary eigenspaces, precluding a single walk operator that encodes the entire trajectory.

8.4.2 The interaction picture for differential equations

A notorious obstacle when applying quantum simulation to spatial PDEs is the stiffness of the Laplacian. In a standard finite-difference discretization of $u_t = -iHu$ with $H = A^{\text{Per}} + V(x)$, the kinetic operator A^{Per} has norm $\|A^{\text{Per}}\| = \alpha_0 = 4/h^2 = \Theta(N^2)$, whereas the potential $V(x)$ has norm $\|V\| = \Theta(1)$. If simulated directly in the Schrödinger picture, the query complexity scales as $O((\alpha_0 + \|V\|)t) = O(N^2 t)$, diverging quadratically with the spatial resolution $N = 1/h$.

The interaction picture simulation method of Low and Wiebe [41, 35] resolves this dilemma whenever the dominant part of H can be fast-forwarded. Write $H = \beta B + \alpha A$, where $\beta = \|B\| \gg \alpha = \|A\|$, and assume $e^{-i\beta B t}$ can be implemented efficiently (for instance, via QFT if B is translation-invariant). Transform the state into the interaction frame:

$$|\psi(t)\rangle_I = e^{i\beta B t} |\psi(t)\rangle. \quad (8.19)$$

Differentiating $|\psi(t)\rangle_I$ using the product rule gives

$$i \frac{d}{dt} |\psi(t)\rangle_I = H_I(t) |\psi(t)\rangle_I, \quad H_I(t) = \alpha e^{i\beta B t} A e^{-i\beta B t}. \quad (8.20)$$

Because unitary conjugation preserves the operator norm,

$$\|H_I(t)\| = \alpha \|A\| \leq \alpha, \quad (8.21)$$

which is *completely independent* of the dominant scale β . The rapid oscillations generated by B are absorbed into the time variation of $H_I(t)$, where $\|H'_I(t)\| \leq \alpha \beta \|[B, A]\|$. In the truncated Dyson

series method, $\|H_I'(t)\|$ enters only logarithmically through the time-register qubit count n_t , while the query complexity depends solely on $\int_0^t \|H_I(s)\| ds \leq \alpha t$.

For the Schrödinger equation with a potential, $B = A^{\text{Per}}$ is diagonalized by the QFT ($B = F^\dagger \Lambda F$), so $e^{-i\beta B t}$ is fast-forwarded in $O(\text{poly}(n))$ gates without querying A . Simulating $H_I(t)$ via the truncated Dyson series then costs only

$$\tilde{O}(\|V\| t \log(1/\epsilon)) \quad (8.22)$$

queries to the potential oracle, entirely eliminating the $O(N^2)$ grid-dependent blow-up.

8.4.3 The no-fast-forwarding theorem and parity reduction

The fast-forwarding of B in the interaction picture is possible only because B possesses special structure (translation invariance). For generic sparse Hamiltonians, time evolution cannot be accelerated: any quantum simulation requires $\Omega(\alpha t)$ queries [8, 35].

The impossibility of generic fast-forwarding is proved by reducing the parity problem to Hamiltonian simulation [8, 35]. Given an unknown n -bit string $s = (s_0, \dots, s_{n-1}) \in \{0, 1\}^n$, deciding the parity $\bigoplus_{j=0}^{n-1} s_j$ requires $\Omega(n)$ queries to the bit oracle $U_s|j\rangle|x\rangle = |j\rangle|x \oplus s_j\rangle$.

Consider the angular momentum operator J_x of spin $n/2$, acting on states $|j\rangle$ for $j \in \{0, \dots, n\}$ with non-zero matrix elements $\langle j+1|J_x|j\rangle = \frac{1}{2}\sqrt{(n-j)(j+1)}$. By the angular momentum algebra $[J_x, J_y] = iJ_z$, the unitary rotation over time $t = \pi$ exactly inverts the z -spin: $e^{-iJ_x\pi}|0\rangle = |n\rangle$. Now embed this rotation into a dual-rail system with an auxiliary parity qubit $|m\rangle$:

$$\langle j+1|\langle m'|G_x|j\rangle|m\rangle = \delta_{m', m \oplus s_j} \frac{1}{2}\sqrt{(n-j)(j+1)}. \quad (8.23)$$

The operator G_x switches rails conditioned on s_j . In a basis permuted by the cumulative parities, G_x is isomorphic to $J_x \otimes I$, which immediately implies

$$e^{-iG_x\pi}|0\rangle|0\rangle = |n\rangle \left| \bigoplus_{j=0}^{n-1} s_j \right\rangle. \quad (8.24)$$

Evolution under G_x for time $t = \pi$ computes the exact parity of the bit string. Because G_x is a 2-sparse banded matrix with $\|G_x\| = \alpha = \Theta(n)$, the dimensionless simulation parameter is $\tau = \alpha t = \Theta(n)$. If a quantum algorithm could simulate generic sparse Hamiltonians in $o(\tau)$ queries, it would compute the parity of s in $o(n)$ queries, contradicting the established quantum query lower bound. Hence, linear scaling $\Omega(\alpha t)$ is a fundamental physical barrier.

8.5 From Hermitian simulation to non-Hermitian generators

8.5.1 The heat semigroup by QSVT

The heat equation $u_t = -Au$ with $A = A^{\text{Per}} \succeq 0$ is not a Hamiltonian simulation, but Section 7.3.3 showed that e^{-tA} is a polynomial of the averaging operator H' of degree $d \approx \sqrt{2z \ln(1/\epsilon)} = \sqrt{\alpha t \ln(1/\epsilon)}$ with $z = \alpha t/2$. The QSVT circuit and the LCU of the even and odd parts give a $(2, m+2)$ -block-encoding of e^{-tA} , with $m = 1$ for the two-term LCU that encodes H' . The cost is $O(\sqrt{\alpha t \log(1/\epsilon)})$ queries, against $\Omega(\alpha t)$ for the unitary evolution. There is no contradiction with the no-fast-forwarding theorem, which concerns unitary evolution. In the Chebyshev expansion (7.14) of the oscillatory filter all coefficients up to $k \approx \tau$ are of the same size, while the coefficients

$e^{-z}I_k(z)$ of the heat filter decay like $e^{-k^2/(2z)}$ from the start: the semigroup damps the high-frequency components, and only the low-degree Chebyshev modes of the generator are needed.

This gain is not free. The block-encoded operator is $e^{-tA}/2$, and the state after postselection is $e^{-tA}|u_0\rangle$ with norm at most $\|u_0\|$. The success probability is $\|e^{-tA}u_0\|^2/(4\|u_0\|^2)$, and amplitude amplification multiplies the query count by the inverse of its square root. For data that decays strongly, such as the lowest Dirichlet mode $e^{-\pi^2 t}$, this overhead grows exponentially with t and can cancel the gain, whereas for data whose mean is preserved on a periodic domain it is bounded. The accounting of this overhead for the heat equation belongs to Chapter 14.

8.5.2 Advection–diffusion

The equation $u_t + cu_x = u_{xx}$ has the generator $A = A^{\text{Per}} + cD_0 = L + iH$ with $L = A^{\text{Per}} \succeq 0$ and $H = -icD_0$ both Hermitian, since $D_0^\dagger = -D_0$. Both are polynomials in the shift S , so they commute and

$$e^{-tA} = e^{-tL} e^{-itH}. \quad (8.25)$$

The right-hand side is the product of the heat-semigroup block of Section 8.5.1 and the transport block of Section 8.3.3, composed by the product rule of Section 6.2. No new algorithm is needed while L and H commute. With variable coefficients, $\nu(x)$ and $c(x)$, they no longer commute, and (8.25) is replaced by a splitting formula of the type of Section 8.1, with e^{-tL} and e^{-itH} as the subroutines and an error governed by $\|[L, H]\|$, or by the constructions below.

8.5.3 Linear combination of Hamiltonian simulation

For a general $A = L + iH$ with $L = (A + A^\dagger)/2 \succeq 0$ and $H = (A - A^\dagger)/(2i)$, linear combination of Hamiltonian simulation (LCHS) [6] writes the semigroup as an integral of Hamiltonian simulations,

$$e^{-tA} = \int_{-\infty}^{\infty} \frac{dk}{\pi(1+k^2)} e^{-it(kL+H)}. \quad (8.26)$$

The integrand is the Hermitian evolution of the generator $kL + H$, and the weight is the Cauchy density. For commuting L and H the identity reduces to the scalar fact $e^{-x} = \int e^{-ikx} dk / (\pi(1+k^2))$ for $x \geq 0$, which is one residue at $k = -i$; the non-commuting case is the theorem of [6]. The weight decays only like k^{-2} , so truncating at $|k| \leq K$ costs about $2/(\pi K)$ and forces $K = O(1/\epsilon)$ with evolution times of order $K\|L\|t$. The first version therefore depends polynomially on $1/\epsilon$, which is why improved weights have been developed (Chapter 10).

8.5.4 Schrödingerization and exact Hermitian dilation

Schrödingerization [28] provides an exact dilation framework embedding general non-Hermitian linear dynamics into unitary Hamiltonian systems via an auxiliary continuous variable.

Consider a general linear differential equation for a system state $|u(t)\rangle$:

$$\frac{d}{dt}|u(t)\rangle = L(t)|u(t)\rangle, \quad (8.27)$$

where $L(t)$ is non-Hermitian, generating dissipation or growth. Decompose $L(t)$ into its skew-Hermitian and Hermitian parts:

$$L(t) = -iH(t) + K(t), \quad H(t) = \frac{i}{2}(L(t) - L(t)^\dagger), \quad K(t) = \frac{1}{2}(L(t) + L(t)^\dagger), \quad (8.28)$$

where both $H(t)$ and $K(t)$ are strictly Hermitian. When $K(t) \neq 0$, evolution under $L(t)$ is non-unitary.

Schrödingerization introduces an auxiliary continuous variable $p \in \mathbb{R}$ with associated Hilbert space $\mathcal{H}_{\text{aux}}$ and canonical Hermitian momentum operator $P_{\text{aux}} = -i\partial_p$.

Theorem (Exact dilation of non-Hermitian evolution) [28]. Define the dilated Hamiltonian $\tilde{H}(t)$ acting on $\mathcal{H}_{\text{sys}} \otimes \mathcal{H}_{\text{aux}}$ by

$$\tilde{H}(t) = H(t) \otimes I_{\text{aux}} + K(t) \otimes P_{\text{aux}}. \quad (8.29)$$

Because $H(t)$, $K(t)$, and P_{aux} are all Hermitian, $\tilde{H}(t)$ is strictly Hermitian. The exact non-Hermitian evolution generated by $L(t)$ is recovered from the dilated unitary evolution by

$$\mathcal{T} \exp\left(\int_0^t L(s) ds\right) = (\langle l| \otimes I_{\text{sys}}) \mathcal{T} \exp\left(-i \int_0^t \tilde{H}(s) ds\right) (|r\rangle \otimes I_{\text{sys}}), \quad (8.30)$$

where $|r\rangle$ and $\langle l|$ are specially chosen auxiliary boundary states. $\square$

Example: Schrödingerization of the heat equation [28]. Consider the classical heat equation $\partial_t u(t, x) = c\nabla^2 u(t, x)$. The spatial operator $L = c\nabla^2 \preceq 0$ is negative semi-definite, so $H = 0$ and $K = L$. Define a warped state $w(t, x, p)$ on $p > 0$ by

$$w(t, x, p) = e^{-p} u(t, x), \quad p > 0. \quad (8.31)$$

Because $\partial_p w = -e^{-p} u = -w$, the dissipative action of L on the system is traded for a derivative with respect to p :

$$\partial_t w = L(e^{-p} u) = -L\partial_p w = -L(iP)w \implies i\partial_t w = (L \otimes P)w. \quad (8.32)$$

Since both L and $P = -i\partial_p$ are Hermitian, the dilated Hamiltonian $\tilde{H} = L \otimes P$ is Hermitian, and the joint state obeys the unitary Schrödinger equation $i\partial_t |w\rangle = \tilde{H}|w\rangle$. After simulating this unitary dynamics on a quantum computer via qubitization or Trotterization, the original classical solution $u(t, x)$ is recovered by integrating out the auxiliary coordinate:

$$\int_0^\infty w(t, x, p) dp = \int_0^\infty e^{-p} u(t, x) dp = u(t, x) \int_0^\infty e^{-p} dp = u(t, x). \quad (8.33)$$

Discretizing p on a finite mesh with a modest number of qubits converts the continuous momentum into a circulant matrix simulatable via QFT or sparse Trotter methods. Schrödingerization thus converts arbitrary dissipative PDEs into standard Hamiltonian simulation, at the cost of one auxiliary spatial register and a postselection/readout step. Further details and higher-order quadratures of LCHS and Schrödingerization are deferred to Chapter 10. Those constructions are quadratures of the same Cauchy identity (Sec. 11.7) that, for Hermitian H , collapses to e^{-iHt} .

Chapter 9

Solving linear systems of equations

The quantum linear systems problem (QLSP) is the linear-algebra core of this report. After discretization (Sec. 5), elliptic PDEs and implicit time steps become instances of $A\mathbf{u} = \mathbf{f}$. A quantum algorithm does not return the classical vector $\mathbf{u}$; it prepares a normalized state encoding $\mathbf{u}$ and leaves observables to be estimated afterwards [25, 35, 34].

Conceptually, every solver in this chapter implements the same scalar map $\lambda \mapsto 1/\lambda$ on the spectrum of A . That map is the Cauchy calculus for $f(z) = 1/z$ (Sec. 11). HHL, QSVT, and a contour quadrature of resolvents are three circuits for that one identity, not three unrelated algorithms. The circuits and costs are developed here; the dictionary that places inversion next to semigroups and Hamiltonian simulation is collected in Sec. 11. None of that is claimed as a new theorem: it is the reading that makes the linear solver the center of the quantum toolkit that follows.

This chapter follows Lin and Wiebe, Chapter 19 [35]: problem setup, the HHL algorithm, the QSVT-based solver, Poisson’s equation as a worked example, and a short outlook on near-optimal adiabatic and quantum-Zeno solvers. Linear combination of Hamiltonian simulation is postponed to Sec. 10; variational solvers are postponed to Sec. 16.

9.1 The quantum linear systems problem

Let $A \in \mathbb{C}^{N \times N}$ be invertible and $|b\rangle$ a normalized state encoding the right-hand side. The QLSP asks for a state

$$|x\rangle = \frac{A^{-1}|b\rangle}{\|A^{-1}|b\rangle\|} \quad (9.1)$$

to accuracy ε in trace distance (or an equivalent state-norm guarantee). Equivalently, if A is accessed through an (α, a, ϵ) -block encoding U_A (Sec. 6), the algorithm is charged for queries to U_A and to a state-preparation unitary U_b with $U_b|0\rangle = |b\rangle$.

Several caveats, emphasized by Li [34], belong in the problem statement rather than in an afterthought:

- the output is a *normalized* quantum state, so the missing factor $\|A^{-1}|b\rangle\|$ must be recovered if a quantity of interest depends on it;
- postselection on the block-encoding ancilla succeeds with a probability that typically scales with $\|A^{-1}|b\rangle\|/\kappa$;
- reading out every amplitude of $|x\rangle$ destroys any asymptotic speedup; the intended output is an observable, an inner product, or a derived PDE quantity.

Throughout, $\kappa = \|A\| \|A^{-1}\|$ denotes the condition number and s the sparsity. For the second-order elliptic matrices of Sec. 5, $\kappa = \Theta(N^{2/d})$ on a d -dimensional grid, which is the dominant parameter in every complexity bound below.

9.2 Classical connection

The discrete Poisson problem

$$A_h \mathbf{u}_h = \mathbf{f}_h \tag{9.2}$$

on a mesh of spacing h is already a QLSP instance: $|b\rangle$ is a normalized encoding of $\mathbf{f}_h$ and $|x\rangle$ encodes $\mathbf{u}_h$. The same pattern appears in every implicit time step of a parabolic scheme (Crank–Nicolson produces a linear system at each step; see Sec. 14) and in the “evolution as a linear system” method for linear differential equations (Sec. 10).

What changes from classical numerical linear algebra is the cost model. Gaussian elimination and Krylov methods return all N entries. A quantum solver’s query complexity may scale polylogarithmically in N , but only after

1. A is block-encoded (cost depends on s and on the rescaling α);
2. $|b\rangle$ is prepared;
3. the algorithm pays a polynomial cost in κ and $1/\varepsilon$;
4. an observable is estimated by sampling.

An apparent speedup in the inversion step need not survive this end-to-end accounting [34]. The discretization chapter planted s , κ , and the mesh-to-dimension map; the rest of this chapter tracks those parameters through concrete algorithms.

9.3 Quantum phase estimation and the HHL algorithm

The Harrow–Hassidim–Lloyd (HHL) algorithm [25] is the historical starting point. In outline, for Hermitian positive-definite A with eigenpairs $(\lambda_j, |v_j\rangle)$ and $|b\rangle = \sum_j \beta_j |v_j\rangle$:

1. apply Hamiltonian simulation of e^{iAt} (Sec. 8) controlled on a clock register;
2. use quantum phase estimation (QPE) to extract an approximation of λ_j into that register;
3. rotate an ancilla by an angle $\arcsin(C/\tilde{\lambda}_j)$ and uncompute the phase register;
4. postselect the ancilla on $|1\rangle$, which (up to normalization) implements $\lambda_j \mapsto C/\lambda_j$ on each eigenspace.

The original complexity scales roughly as $\tilde{O}(\kappa^2 s/\varepsilon)$ up to polylogarithmic factors in N , reflecting QPE’s precision requirements and amplitude amplification over the postselection probability [25, 35]. Subsequent work improved the κ and ε exponents; those improvements are the subject of the QSVT and adiabatic/Zeno constructions below.

For this report HHL is kept short: it shows that inversion is a controlled eigenvalue transformation plus postselection, and it makes QPE available as a tool without a separate eigenvalue chapter. Circuit-level error analysis is omitted; the conceptual map

$$|b\rangle \mapsto \sum_j \beta_j |v_j\rangle \mapsto \sum_j \frac{\beta_j}{\lambda_j} |v_j\rangle \propto A^{-1} |b\rangle \tag{9.3}$$

is what later PDE chapters reuse.

9.4 QSVT-based linear-system solver

Modern fault-tolerant solvers replace QPE by a polynomial approximation of $1/x$ applied directly to a block encoding [23, 35, 34]. On the spectral window $[-1, -1/\kappa] \cup [1/\kappa, 1]$ one constructs a bounded polynomial p of degree $\tilde{O}(\kappa \log(1/\varepsilon))$ such that

$$|p(x) - 1/x| \leq \varepsilon, \quad x \in [-1, -1/\kappa] \cup [1/\kappa, 1]. \quad (9.4)$$

QSVT (Sec. 7) implements $p(A/\alpha)$ as a block of a larger unitary. Postselection on the ancilla then prepares a state close to (9.1).

Compared with HHL, the QSVT route

- avoids an extra clock register and the associated QPE precision overhead;
- treats non-Hermitian A through singular-value transformation (inversion becomes a function of $A^\dagger A$, or a Hermitian dilation);
- matches the “inverse filter” already introduced as a QSVT example in Li, Sec. 1.12.1 [34].

Lin’s Theorem 19.3 [35] is the reference complexity statement to be quoted when this subsection is expanded. Linear-combination-of-unitaries constructions of the same filter (Lin, Exercise 19.1) are noted but not developed.

The PDE-relevant special case is a Hermitian positive-definite discretization A_h with known spectral bounds. Then the only polynomial needed is an odd approximation to $1/x$ on $[1/\kappa, 1]$, and the block encoding can be the sparse-matrix encoding of the stencil from Sec. 6.

9.5 Cauchy’s formula as the common writing

The three inversion methods used in this report are the same holomorphic map in different coordinates [50, 27, 23]:

$$A^{-1} = \frac{1}{2\pi i} \oint_{\Gamma} z^{-1} (zI - A)^{-1} dz. \quad (9.5)$$

HHL evaluates $1/\lambda_j$ after reading eigenvalues (a residue/spectral form of (9.5)). QSVT replaces $1/z$ by a polynomial on the real spectral window. A contour algorithm quadrature-discretizes (9.5) and calls a QLSP at each node z [50, 51]. The linear solver is therefore not next to Cauchy’s theorem; it is the $f(z) = 1/z$ case, with the resolvent as the integrand. Semigroups and Hamiltonian simulation are the neighboring cases e^{tz} and e^{-itz} , postponed to Secs. 10 and 11.

9.6 Worked example: Poisson’s equation

Consider the model Poisson problem on the unit cube (periodic or homogeneous Dirichlet), discretized by the standard $2d + 1$ -point Laplacian A_h . The pipeline of this report becomes fully concrete:

$$-\Delta u = f \longrightarrow A_h \mathbf{u}_h = \mathbf{f}_h \longrightarrow \text{block encoding of } A_h/\alpha \longrightarrow \text{QSVT inverse filter} \longrightarrow |u_h\rangle. \quad (9.6)$$

Items to fill in when the section is written out:

- the exact stencil, the scaling $\alpha = \Theta(h^{-2})$, and $\kappa(A_h) = \Theta(h^{-2})$;

- preparation of $|f_h\rangle$ for a smooth right-hand side (Grover–Rudolph / QFT diagonalization when the grid is periodic);
- the distinction between preparing $|u_h\rangle$ and estimating a linear functional $\mathbf{c}^\top \mathbf{u}_h$;
- the accuracy-matched comparison with a classical multigrid solve as a function of d , with and without preconditioning, deferred from Section 1.5;
- a pointer forward to Sec. 12, which reuses this example with first-order factorization, Hermitian dilation, and FEM mass matrices rather than repeating the QLSP reduction.

Green’s functions of fermionic systems (Lin Sec. 19.5) are omitted: they are not needed for the PDE track.

9.7 Near-optimal and optimal solvers

HHL and the basic QSVT inverter are not complexity-optimal in κ . Two constructions from Lin, Secs. 19.6–19.7 [35], are recorded here as an outlook, not as algorithms this report will implement:

- *adiabatic near-optimal solvers*, which encode inversion as a slow eigenpath of a family of Hamiltonians and remove a κ factor relative to textbook HHL [47, 5];
- *quantum-Zeno optimal solvers*, which replace continuous adiabatic evolution by a discrete sequence of projections and achieve the optimal $\Theta(\kappa \log(1/\varepsilon))$ query scaling up to logarithmic factors [35].

For PDE applications the leading practical messages are already in Secs. 9.2–9.6: control κ by the discretization, encode the stencil, invert by QSVT, and measure an observable. The adiabatic/Zeno literature is the complexity ceiling, not the pedagogical entry point.

9.8 What this chapter does not cover

- LCHS, time marching, and the $f(z) = e^{tz}$ side of the same calculus — Secs. 10 and 11.7;
- VQLS and other NISQ heuristics — Sec. 16;
- classical rootfinding for nonlinear algebraic systems — Sec. 15;
- the derivation of QSP phase factors — Sec. 7.

Chapter 10

Solving linear differential equations

Linear differential equations

$$\mathbf{u}'(t) = A\mathbf{u}(t) + \mathbf{f}(t) \tag{10.1}$$

are the time-dependent counterpart of the QLSP. Three standard quantum routes, following Lin, Chapter 20 [35], are:

- rewrite the evolution as one large linear system and call a QLSP solver (Sec. 9);
- time marching, i.e. a sequence of short QLSP or simulation steps;
- linear combination of Hamiltonian simulation (LCHS), which expresses a non-Hermitian semigroup as an integral (or quadrature) of Hermitian simulations [6, 34].

10.1 Problem setup

10.2 The linear-systems method

10.3 Time marching

10.4 Linear combination of Hamiltonian simulation

LCHS lives here, not in the QSP/QSVT chapter. Schrödingerization [28] is noted as a closely related dilation and is reused for hyperbolic problems in Sec. 13.

10.5 Example: the heat equation

Forward pointer to Sec. 14 (Crank–Nicolson as QLSP; QSVT heat semigroup).

10.6 Non-Hermitian generators and complex absorbing potentials

The contour, Fourier, and dilation pictures that make LCHS and Schrödingerization the same task as “apply e^{tA} ” are collected in Sec. 11.7.

Chapter 11

Complex analysis as a language for PDEs and quantum algorithms

Lin and Li develop the quantum algorithms as a toolkit: block encodings, polynomial transformations, Hamiltonian simulation, linear-system solvers, and LCHS appear as successive primitives [35, 34]. This chapter only rearranges that toolkit around one identity, the Cauchy integral formula lifted to operators. The linear solver of Sec. 9 is the case $f(z) = 1/z$; Hamiltonian simulation and LCHS are neighboring choices of f . The algorithms are standard [50, 6, 27]; the point here is to write them so that Cauchy sits at the center of the solver, which is the picture this report wanted to make visible.

11.1 The identity

The scalar Cauchy formula [2]

$$f(a) = \frac{1}{2\pi i} \oint_{\Gamma} \frac{f(z)}{z - a} dz \quad (11.1)$$

says that a holomorphic function is determined by its values on a contour around the point. For a matrix or closed operator A whose spectrum lies inside a positively oriented contour Γ , the same writing is the holomorphic functional calculus [26]

$$f(A) = \frac{1}{2\pi i} \oint_{\Gamma} f(z) (zI - A)^{-1} dz. \quad (11.2)$$

Two elementary observations turn (11.2) into a dictionary for this report.

1. The integrand contains the resolvent $R(z; A) = (zI - A)^{-1}$. Applying $R(z; A)$ to a vector is a linear system, i.e. a QLSP instance (Sec. 9).
2. A quadrature of (11.2) is a linear combination of resolvents, hence a linear combination of unitaries once each resolvent (or each Hermitian evolution that represents it) is block-encoded. That is LCU. On a real spectrum the integral may be replaced by a polynomial in A , which is QSVT.

So Cauchy's theorem is not being identified with quantum computing as a model of computation. It is being used as the *analysis* of what the computation is doing: evaluating f on the spectrum of A by probing a neighborhood of that spectrum.

11.2 Three levels of the interpretation

It helps to separate what is being claimed.

- *Task equivalence.* Hamiltonian simulation, QLSP, heat semigroups, and spectral filters are the same problem—prepare a state proportional to $f(A)|b\rangle$ —for different holomorphic (or uniformly approximable) f . This is already implicit in QSVT [23]; the Cauchy calculus makes the non-Hermitian case equally visible.
- *Representation equivalence.* Equation (11.2), a Fourier/Laplace inversion, and a polynomial $p \approx f$ on $\text{spec}(A)$ are three writings of the same map $A \mapsto f(A)$. LCHS is a quadrature of the first or second writing; QSVT is the third. Residue calculus is the spectral-projection case $f \equiv 1$ on a component of the spectrum.
- *Not a complexity equivalence.* The contour does not cancel κ , state-preparation cost, or measurement (Sec. 5). A contour that hugs $\text{spec}(A)$ reproduces the same blow-up of the resolvent that HHL pays in κ . The interpretation organizes algorithms; it does not dominate their resource counts.

The first two levels are the organizing picture used in this report. Lin invokes contours for non-Hermitian dynamics and complex absorbing potentials [35]; Li uses inverse and exponential filters as QSVT examples [34]. Neither notes volume is arranged as a Cauchy commentary on the PDE pipeline, which is why the identity is written out here. The algorithms themselves are not new. Section 11.3 cites the line that already implements Cauchy’s formula on a quantum computer, and the parallel line that implements the same maps by Hamiltonian simulation.

11.3 What is already known

Cauchy’s integral formula as a *quantum algorithm* for $f(A)|b\rangle$ begins with Takahira, Ohashi, Sogabe, and Usuda [50]: a circular contour, trapezoidal quadrature, and a linear combination of shifted inverses, with polylogarithmic dependence on the error and without requiring A to be Hermitian. The same group recast the construction in the block-encoding framework and allowed a general contour [51]. Jiang and An [27] then gave a complete complexity analysis of that contour method and applied it to Hamiltonian simulation ($f(z) = e^{Tz}$, $A = -iH$), matrix polynomials, and linear ODEs. For Hermitian Hamiltonian simulation the contour route is *not* optimal: they obtain $\tilde{O}(T^2 \log(1/\varepsilon))$, against the qubitization/QSVT scaling $O(T + \log(1/\varepsilon))$ [40, 23]. Where the contour method can win is asymptotically stable linear differential equations, not unitary evolution.

A second family never forms the resolvent explicitly. Linear combination of Hamiltonian simulation (LCHS) writes a dissipative semigroup as an integral of unitary evolutions $e^{-it(H+kL)}$ [6]. Laplace-LCHS extends the same idea from e^{-AT} to a class of eigenvalue transformations that admit an inverse Laplace transform, including A^{-k} and $e^{-A^{-1}}$, still using only Hamiltonian simulation as the primitive [3]. Rational transformations via integral representations of the resolvent, implemented by LCU of Hamiltonian simulations, sit in between the two families [45]. Schrödingerization is the dilation cousin: an extra momentum coordinate turns non-unitary dynamics into a Schrödinger equation [28].

The schematic is therefore already in the literature:

contour + QLSP (resolvents) versus integral kernel + Hamiltonian simulation
(LCHS),

with QSVT as the real-spectrum polynomial surrogate. Hamiltonian simulation is the *engine* of the second family and a *test case* of the first; it is not an undiscovered equivalent of Cauchy’s theorem.

11.4 A dictionary

Complex analysis	Linear algebra / PDE	Quantum primitive
point a ; function f	spectrum of A ; $f(A)$	block encoding of A , then f
$1/(z - a)$	resolvent $(zI - A)^{-1}$	QLSP at the shift z
Cauchy integral (11.2)	holomorphic calculus	LCU quadrature of resolvents
polynomial approximation of f on a cut	$p(A)$ on a real spectrum	QSP / QSVT
$f(z) = 1/z$	linear solve $A\mathbf{u} = \mathbf{b}$	HHL / QSVT inverter
$f(z) = e^{-itz}$	unitary group, hyperbolic/Schrödinger	Hamiltonian simulation
$f(z) = e^{tz}$	semigroup, parabolic generator	QSVT heat filter or LCHS
residue at a pole	eigenprojection, Green’s residues	QPE; QSVT step/sign filter
argument principle	eigenvalue counting inside Γ	contour of resolvents / QPE
contour deformation	analytic continuation of e^{tA}	LCHS kernels; CAP

Each row is a translation, not a reduction of one column to another. The middle column is the PDE and discretization layer; the right column is the quantum toolkit; the left column is the analysis that makes the two speak.

11.5 From Cauchy–Riemann to elliptic PDEs

Chapter 4 proves the scalar theory: Cauchy’s theorem, the equivalent characterizations of holomorphic functions, the harmonic conjugate, the Poisson kernel of the disk as the real part of the Cauchy kernel, and the conformal invariance of Δ in the plane. The operator-level statement used below is the same integral with the scalar kernel $1/(z - a)$ replaced by the resolvent. Elliptic inversion and Green’s functions are contour and residue statements about $(\zeta I - \Delta)^{-1}$.

11.6 Resolvents, Green’s functions, and linear systems

For each $z \notin \text{spec}(A)$, applying $R(z; A)$ to a vector is a QLSP. Cauchy’s formula does not replace the system solver; it *calls* it at quadrature nodes and superposes the outputs. Specializing $f(z) = 1/z$ (with 0 outside Γ , or with Γ split around a gap) recovers

$$A^{-1} = \frac{1}{2\pi i} \oint_{\Gamma} z^{-1} (zI - A)^{-1} dz, \quad (11.3)$$

the same inversion that HHL and QSVT implement by a real-line filter. The Green’s function of an elliptic operator is the integral kernel of a resolvent; the formulas for those kernels are in Chapter 3. Residues of $R(z; A)$ are eigenprojections. Poles encode eigenvalues, a contour that avoids the spectrum encodes a matrix function, and a contour that collapses onto a cut encodes a semigroup.

11.7 Semigroups, Hamiltonian simulation, and LCHS

Three choices of f recover the rest of the quantum toolkit.

- $f(z) = e^{-itz}$ with Hermitian $A = H$ and spectrum on $\mathbb{R}$: Hamiltonian simulation (Sec. 8). The contour deforms onto the real line; QSP/QSVT replace the integral by a polynomial in H .
- $f(z) = e^{tz}$ with a possibly non-Hermitian generator A : the solution operator of $\mathbf{u}' = A\mathbf{u}$ (Sec. 10). Analytic semigroups for parabolic problems are the same statement with a sectorial contour [35, 34].
- A Fourier or Laplace writing of e^{tA} as an integral of Hermitian evolutions e^{-iHs} is LCHS [6]: quadrature of (11.2) (or of its Fourier cousin) implemented as LCU of Hamiltonian simulations. Schrödingerization [28] dilates a non-unitary generator into a Schrödinger equation—again a contour/Fourier inversion realized by unitary dynamics.

Complex absorbing potentials [35] shift the contour: A is perturbed into the complex plane so that a resonance or an outgoing radiation condition becomes a resolvent problem on the deformed path. That is contour deformation as a PDE boundary condition, not as an optional remark.

11.8 Residues, projectors, and what QPE is counting

The residue of the resolvent at an isolated eigenvalue λ is the spectral projection onto the corresponding invariant subspace. In the scalar theory this is Cauchy’s residue theorem; in the operator theory it is the Riesz projection

$$P_\lambda = \frac{1}{2\pi i} \oint_{|z-\lambda|=\varepsilon} (zI - A)^{-1} dz. \quad (11.4)$$

Quantum phase estimation samples eigenvalues of a unitary; QSVT step and sign filters build approximate spectral projections without an extra clock register. Both are quantum writings of P_λ . The argument principle, which counts zeros inside Γ by a winding number, likewise has a linear algebra twin: the number of eigenvalues of A inside Γ is $\text{tr } P_\Gamma$. A quantum algorithm that estimates that trace is counting spectrum by a contour of resolvents. This row of the dictionary is the natural home of the eigenvalue week that QISCA listed separately: it is not a third primitive besides QLSP and simulation, but the residue case of the same calculus.

11.9 QSVT as a real-spectrum surrogate

QSVT does not evaluate a contour integral. It implements a polynomial $p(A)$ (or of the singular values) given a block encoding [23]. On the spectrum of a Hermitian or normal matrix this agrees with (11.2), because a holomorphic f may be approximated uniformly by polynomials on a real interval—the inverse, heat, and oscillatory filters of Sec. 7. The working slogan of the report is therefore

holomorphic functional calculus: Cauchy integral of resolvents;
 QSVT: polynomial calculus on the (singular) spectrum;
 LCHS: quadrature of a Cauchy/Fourier identity by Hamiltonian simulation;
 residue / Riesz projection: spectral filtering and QPE.

Same task $A \mapsto f(A)$; three circuits; one analysis.

Part III

Quantum Algorithms for Differential Equations

Chapter 12

Quantum algorithms for elliptic PDEs

Elliptic problems are the direct application of the QLSP. After the discretization $A_h \mathbf{u}_h = \mathbf{f}_h$, one applies the solver of Sec. 9 and estimates a quantity of interest. In two dimensions the same inversion is Laplace’s equation for a holomorphic potential (Sec. 11.5); the quantum chapter does not need conformal maps, only the resolvent.

12.1 Classical discretization and the direct QLSP route

Reuses the Poisson example of Sec. 9.6.

12.2 First-order factorization and Hermitian dilation

12.3 QFT-based spectral filtering

12.4 Elliptic eigenvalue problems and phase estimation

Chapter 13

Quantum algorithms for hyperbolic PDEs

Hyperbolic problems (waves, first-order transport after discretization) map primarily onto Hamiltonian simulation and LCHS rather than onto a single linear solve. The classical reduction is a first-order system, optionally dilated to a Schrödinger equation [34, 28].

13.1 Classical wave discretizations

13.2 From the wave equation to a Schrödinger equation

13.3 Schrödingerization

13.4 LCHS for non-Hermitian and forced hyperbolic systems

Reuses Sec. 10.

Chapter 14

Quantum algorithms for parabolic PDEs

Parabolic problems (heat, diffusion) sit between the QLSP and Hamiltonian simulation. Two standard quantum routes [34, 35]:

- Crank–Nicolson (or another implicit scheme) as a sequence of linear systems, each solved by Sec. 9;
- QSVT implementation of the heat semigroup e^{tA} (the exponential filter of Sec. 7).

14.1 Classical heat discretizations

14.2 Crank–Nicolson as a quantum linear system

14.3 QSVT implementation of the heat semigroup

14.4 Hermitian dilation and Gaussian transmutation

Chapter 15

Nonlinear PDEs

Nonlinear equations do not yield a single sparse linear system. Section 3.6 records the classical facts used below: the shock speed (3.161), the entropy condition (3.162), the KdV soliton (3.168), the acoustic limit (3.173), and the incompressible Navier–Stokes system (3.174). The quantum route used in this report is to *lift* a nonlinear ODE/PDE to a larger linear dynamics (Carleman linearization, Kolmogorov/Liouville, or Koopman–von Neumann formulations) and then apply the linear tools of Secs. 9–10 [34]. The classical iterations are Section 5.2. Each Newton step on a discretized nonlinear system is again a linear system, hence a QLSP instance.

15.1 Why nonlinear PDEs require a different viewpoint

15.2 From nonlinear PDEs to nonlinear ODEs

15.3 Carleman linearization

15.4 Kolmogorov, Liouville, and Koopman–von Neumann formulations

15.5 Classical rootfinding as a supporting tool

Section 5.2 records the scalar iterations and the Jacobian step (5.37). For a nonlinear discretization the unknown in that step is the grid vector, and $F(x_n)$ is a sparse matrix of the kind assembled in Chapter 5.

Part IV

NISQ Implementations and Research Frontiers

Chapter 16

Variational linear solvers and NISQ methods

Variational quantum linear solvers (VQLS) minimize a residual-based cost function on a shallow parameterized circuit [12]. They are a NISQ heuristic, not a fault-tolerant replacement for HHL or QSVT, and are kept out of the QSP/QSVT and QLSP chapters for that reason.

16.1 The VQLS cost function

16.2 Relation to the QLSP

16.3 Limitations and when a variational method is not a solver

Chapter 17

Latest research and outlook

Reserved for the QISCA closing weeks: recent quantum PDE algorithms, software stacks (block-encoding toolboxes, Schrödingerization workflows), and an end-to-end assessment of where a genuine advantage could appear after discretization error, state preparation, and measurement are counted.

17.1 Recent algorithmic developments

17.2 Software and programming interfaces

17.3 Open questions

Further reading, not a new claim: for a given discretized PDE, compare the Cauchy–QLSP contour route [27], LCHS [6, 3], and QSVT on the same matrix, including κ , quadrature size, and measurement. Jiang and An already note a possible win for asymptotically stable ODEs. This report stops at arranging those algorithms around $f(z) = 1/z$ at the linear solver.

Bibliography

- [1] Scott Aaronson. Read the fine print. *Nature Phys.*, 11:291–293, 2015.
- [2] Lars V. Ahlfors. *Complex Analysis*. McGraw-Hill, New York, 3rd edition, 1979.
- [3] Dong An, Andrew M. Childs, Lin Lin, and Lexing Ying. Laplace transform-based quantum eigenvalue transformation via linear combination of Hamiltonian simulation. *SIAM J. Comput.*, 55(2), 2026.
- [4] Dong An, Di Fang, and Lin Lin. Time-dependent unbounded Hamiltonian simulation with vector norm scaling. *Quantum*, 5:459, 2021.
- [5] Dong An and Lin Lin. Quantum linear system solver based on time-optimal adiabatic quantum computing and quantum approximate optimization algorithm. *ACM Trans. Quantum Comput.*, 3(2):5, 2022.
- [6] Dong An, Jin-Peng Liu, and Lin Lin. Linear combination of Hamiltonian simulation for nonunitary dynamics with optimal state preparation cost. *Phys. Rev. Lett.*, 131:150603, 2023.
- [7] Kendall E. Atkinson. *An Introduction to Numerical Analysis*. John Wiley & Sons, New York, 2 edition, 1989.
- [8] Dominic W. Berry, Graeme Ahokas, Richard Cleve, and Barry C. Sanders. Efficient quantum algorithms for simulating sparse Hamiltonians. *Commun. Math. Phys.*, 270:359–371, 2007.
- [9] Dominic W. Berry, Andrew M. Childs, Richard Cleve, Robin Kothari, and Rolando D. Somma. Simulating Hamiltonian dynamics with a truncated Taylor series. *Phys. Rev. Lett.*, 114:090502, 2015.
- [10] Dominic W. Berry, Andrew M. Childs, and Robin Kothari. Hamiltonian simulation with nearly optimal dependence on all parameters. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, pages 792–809, 2015.
- [11] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *J. Polit. Econ.*, 81(3):637–654, 1973.
- [12] Carlos Bravo-Prieto, Ryan LaRose, M. Cerezo, Yigit Subasi, Lukasz Cincio, and Patrick J. Coles. Variational quantum linear solver. *Quantum*, 7:1188, 2023.
- [13] Daan Camps, Lin Lin, Roel Van Beeumen, and Chao Yang. Explicit quantum circuits for block encodings of certain sparse matrices. *SIAM Journal on Matrix Analysis and Applications*, 45(1):801–827, 2024.

- [14] Andrew M. Childs, Robin Kothari, and Rolando D. Somma. Quantum algorithm for systems of linear equations with exponentially improved dependence on precision. *SIAM J. Comput.*, 46(6):1920–1950, 2017.
- [15] Andrew M. Childs, Yuan Su, Minh C. Tran, Nathan Wiebe, and Shuchen Zhu. Theory of Trotter error with commutator scaling. *Phys. Rev. X*, 11:011020, 2021.
- [16] Andrew M. Childs and Nathan Wiebe. Hamiltonian simulation using linear combinations of unitary operations. *Quantum Information and Computation*, 12(11–12):901–924, 2012.
- [17] Robert Davis. How IBM is searching for quantum advantage in differential equations. IBM Research Blog, August 2026. Q&A with Hari Krovi.
- [18] Yulong Dong, Lin Lin, Hongkang Ni, and Jiasu Wang. Infinite quantum signal processing. *Quantum*, 8:1558, 2024.
- [19] Yulong Dong, Lin Lin, Hongkang Ni, and Jiasu Wang. Robust iterative method for symmetric quantum signal processing in all parameter regimes. *SIAM J. Sci. Comput.*, 46(4):A2951–A2971, 2024.
- [20] Yulong Dong, Xiang Meng, K. Birgitta Whaley, and Lin Lin. Efficient phase-factor evaluation in quantum signal processing. *Phys. Rev. A*, 103:042419, 2021.
- [21] Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, 2 edition, 2010.
- [22] Richard P. Feynman. Simulating physics with computers. *Int. J. Theor. Phys.*, 21:467–488, 1982.
- [23] András Gilyén, Yuan Su, Guang Hao Low, and Nathan Wiebe. Quantum singular value transformation and beyond: Exponential improvements for quantum matrix arithmetics. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 193–204, 2019.
- [24] Jeongwan Haah. Product decomposition of periodic functions in quantum signal processing. *Quantum*, 3:190, 2019.
- [25] Aram W. Harrow, Avinandan Hassidim, and Seth Lloyd. Quantum algorithm for linear systems of equations. *Phys. Rev. Lett.*, 103:150502, 2009.
- [26] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, Philadelphia, 2008.
- [27] Shan Jiang and Dong An. Contour-integral based quantum eigenvalue transformation: Analysis and applications. arXiv:2601.11959, 2026.
- [28] Shi Jin, Nana Liu, and Yue Yu. Quantum simulation of partial differential equations: Applications and detailed analysis. *Phys. Rev. A*, 108:032603, 2023.
- [29] Jeongbin Jo. Numerical partial differential equations: Advanced methods and analysis. Study notes, Yonsei University, March 2026.
- [30] Kyungkeun Kang. Summary of partial differential equations. Lecture notes, Yonsei University, 2025.

- [31] Ryogo Kubo. Statistical-mechanical theory of irreversible processes. I. general theory and simple applications to magnetic and conduction problems. *Journal of the Physical Society of Japan*, 12(6):570–586, 1957.
- [32] E. Lee. Numerical analysis. Lecture notes, Computational Science and Engineering, Yonsei University. The initial-value theory follows Atkinson, *An Introduction to Numerical Analysis*, 2nd ed., Wiley, 1989, 2026.
- [33] Randall J. LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*. SIAM, Philadelphia, 2007.
- [34] Xiantao Li. A quantum path to partial differential equations. Version 1.0, July 2026.
- [35] Lin Lin and Nathan Wiebe. Quantum algorithms for scientific computation. Live lecture notes, 29 April 2026, April 2026.
- [36] Noah Linden, Ashley Montanaro, and Changpeng Shao. Quantum vs. classical algorithms for solving the heat equation. *Commun. Math. Phys.*, 395:601–641, 2022.
- [37] Jin-Peng Liu, Herman Øie Kolden, Hari K. Krovi, Nuno F. Loureiro, Konstantina Trivisa, and Andrew M. Childs. Efficient quantum algorithm for dissipative nonlinear differential equations. *Proc. Natl. Acad. Sci. USA*, 118(35):e2026805118, 2021.
- [38] Seth Lloyd. Universal quantum simulators. *Science*, 273:1073–1078, 1996.
- [39] Guang Hao Low and Isaac L. Chuang. Optimal Hamiltonian simulation by quantum signal processing. *Phys. Rev. Lett.*, 118:010501, 2017.
- [40] Guang Hao Low and Isaac L. Chuang. Hamiltonian simulation by qubitization. *Quantum*, 3:163, 2019.
- [41] Guang Hao Low and Nathan Wiebe. Hamiltonian simulation in the interaction picture. *Phys. Rev. Lett.*, 125:040502, 2019.
- [42] Ashley Montanaro and Sam Pallister. Quantum algorithms and the finite element method. *Phys. Rev. A*, 93:032324, 2016.
- [43] NIST. NIST digital library of mathematical functions. <https://dlmf.nist.gov/>. Sections 10.12, 10.14 and 10.35.
- [44] Hannes Risken. *The Fokker–Planck Equation: Methods of Solution and Applications*. Springer, Berlin, 2 edition, 1989.
- [45] Yizhi Shen, Niel Van Buggenhout, Daan Camps, Katherine Klymko, and Roel Van Beeumen. Quantum rational transformation using linear combinations of Hamiltonian simulations. arXiv:2408.07742, 2024.
- [46] Elias M. Stein and Rami Shakarchi. *Complex Analysis*, volume 2 of *Princeton Lectures in Analysis*. Princeton University Press, Princeton, 2003.
- [47] Yiğit Subaşı, Rolando D. Somma, and Davide Orsucci. Quantum algorithms for systems of linear equations inspired by adiabatic quantum computing. *Phys. Rev. Lett.*, 122:060504, 2019.

- [48] Masuo Suzuki. Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations. *Phys. Lett. A*, 146:319–323, 1990.
- [49] Masuo Suzuki. General theory of fractal path integrals with applications to many-body theories and statistical physics. *J. Math. Phys.*, 32:400–407, 1991.
- [50] Souichi Takahira, Asuka Ohashi, Tomohiro Sogabe, and Tsuyoshi Sasaki Usuda. Quantum algorithm for matrix functions by Cauchy’s integral formula. *Quantum Inf. Comput.*, 20(1–2):14–36, 2020.
- [51] Souichi Takahira, Asuka Ohashi, Tomohiro Sogabe, and Tsuyoshi Sasaki Usuda. Quantum algorithms based on the block-encoding framework for matrix functions by contour integrals. *Quantum Inf. Comput.*, 22(11–12):965–979, 2021.
- [52] Lloyd N. Trefethen. *Approximation Theory and Approximation Practice*. SIAM, Philadelphia, 2013.
- [53] H. F. Trotter. On the product of semi-groups of operators. *Proc. Amer. Math. Soc.*, 10:545–551, 1959.
- [54] Jiasu Wang, Yulong Dong, and Lin Lin. On the energy landscape of symmetric quantum signal processing. *Quantum*, 6:850, 2022.
- [55] Haruo Yoshida. Construction of higher order symplectic integrators. *Phys. Lett. A*, 150:262–268, 1990.