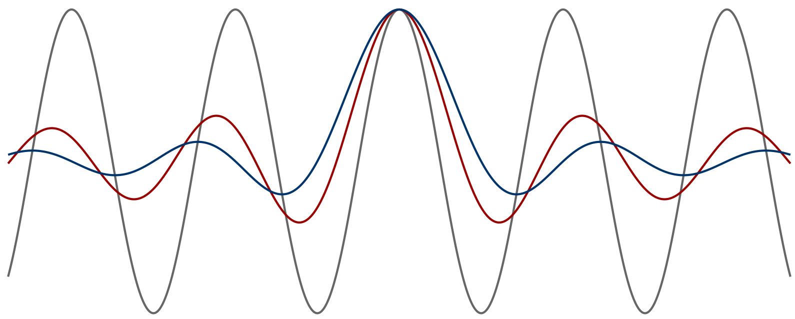
Geometrical Shape as a Cause

Orthogonal Decompositions and Mellin Transform Methods



David Bakker

2026

Master's Thesis

Submitted in partial fulfillment of the requirements for the degrees in

Mathematical Sciences
and
History and Philosophy of Science

Utrecht University

Geometrical Shape as a Cause

Orthogonal Decompositions and Mellin Transform Methods

Supervisors

Viktor Blåsjö
Michał Wrochna

Mathematical Institute, Utrecht University

Author

David Bakker

July 2026

Cover illustration: A contour plot of the cosine, spherical Bessel and Bessel function $J_0(x)$, arising in Fourier's analysis of heat flow in a cube, sphere, and cylinder, respectively. See Chapter 2 section II.

Contents

Introduction

This thesis is the final project for two masters: Mathematics and History & Philosophy of Science. It therefore contains a substantial mathematical as well as historical and philosophical component. My goal was not merely to satisfy this as a formal requirement, but to write a thesis that combines both as part of a single thesis united by a philosophical theme spanning across the historical and mathematical components. Three additional subthemes are confined to parts of the thesis. In this introduction I will say something briefly about all and subsequently indicate what are novel ideas in the thesis.

Geometrical shape as a cause

In the *Posterior Analytics*, Aristotle makes a distinction between a deduction and a demonstration. A deduction establishes a conclusion by logical inference from given premises; a demonstration is a type of deduction that produces scientific knowledge—*epistēmē*. For a demonstration it is not sufficient that the premises of the deduction are true and that the argument is formally valid, the premises have to be of a certain kind in order to produce scientific knowledge. For example, consider the following deduction:

All planets are near the Earth.
Everything near the Earth does not twinkle.

∴ Planets do not twinkle.

This is a valid deduction, but not a demonstration, since being near the earth is not the cause of not twinkling. In gaining epistēmē we are moving from knowing *that* certain things are so to knowing *why* they are so. A demonstration must capture *why* the phenomenon occurs, it must exhibit its cause, not merely derive its truth. Aristotle applies this view of scientific knowledge to mathematics too: plane geometry is his recurring example of a genuine science that produces epistēmē. In particular consider the following example of epistēmē in mathematics:

> For it is for the doctor to know the fact that circular wounds heal more slowly, and for the geometer to know the reason why.

Aristotle 1984, APo I.13, 79a. What I take him to have in mind is that wounds heal at their perimeter and that a circle is the shape with the smallest perimeter for a given area.

The fact has a geometrical cause. This thesis takes the idea that geometrical facts can be causes and applies it to cases within an important mathematical development: spectral analysis, or, the decomposition of functions into special functions.

Within the broader scientific enterprise of looking for demonstrations, Aristotle gives a perspective that four things that can serve as genuine middle terms in such demonstrations: the efficient, material, formal and final cause. Geometrical shape would be an instance of a formal cause. Aristotle's doctrine of the four causes was largely repudiated in the early modern period and my goal will not be to revive it. However, what was rejected was *not* the general idea of providing middle terms as why's, but a particular vacuous manner of giving demonstrations: why does opium induce sleep? Because it possesses a dormitive virtue. Such answers—in terms of substantial forms and the occult qualities held to flow from them—were rejected for being empty causes, they 'explain' a phenomenon by renaming it. Such arguments do not supply a genuine middle term and therefore even fail Aristotle's own standards. A question that remains is what may legitimately be a middle term, in particular, what the middle terms should be in a *mathematical* science of nature.

One kind of answer was provided by Laplace: 'the phenomena of nature are in the last analysis reduced to actions *ad distans* from molecule to molecule, and that the consideration of these actions must serve as the basis of the mathematical theory of these phenomena.' Laplace encourages us to seek middle terms that reduce to particles and molecular forces between them. In developing a mathematical theory of heat, Fourier rejects this Laplacian program that puts molecular processes as the middle term. He gives, first of all, a reformulation of what the nature of the mathematical problem of the evolution of temperatures is: an initial boundary value problem, rather than the solution of a PDE on an unbounded domain. The geometric domain thus gets a central place in what the mathematical formulation of the phenomenon—the movement of heat—is. We find in Fourier the idea that temperature distributions have to be broken down into 'simple modes' and that these simple modes govern the behavior of heat. It is the shape of the domain that determines the modes, while the initial data only determines how much of each is present. In the movement to equilibrium the initial states are forgotten—the non-equilibrium modes decay away, and the steady distribution that remains is fixed by the domain and the conditions at its boundary, with no relation to the initial state. If these facts give us epistēmē about the movement of heat, then the proper middle terms are derived from the geometrical shape, not from molecular forces. The idea of decomposing a function into special functions is not original to Fourier; Laplace, too, decomposes functions into orthogonal parts, e.g. the spherical harmonics of the potential. However, Laplace does not attach the physical significance to it that Fourier does, it remains a mathematical technique for solving the equation.

See p. 24 for the context of this quote.

By mentioning, but not reviving, the Aristotelian notion of a formal cause, I hope to separate this notion from the broader idea of giving middle terms that are causes and lead to epistēmē. Spectral analysis has been central to mathematics and mathematical physics from Fourier's work onwards. While the spectral perspective and its physical interpretation is found in Fourier, the Aristotelian framing is my own and no explicit references to Aristotle are found in Fourier's work on the mathematics of heat. Fourier is normally framed as *rejecting* a kind of explanation offered by the Laplacians, through this perspective I hope to give a positive account of what insight his work provides.

This theme extends to the purely mathematical Chapter 4, where we consider an advanced modern instance of the spectral perspective. This chapter focusses on the Mellin transform, a close relative of the Fourier transform. Its use allows us to bring same kind of causal perspective found in Chapter 2 on heat to examples of decompositions of Laplace discussed in Chapter 1, in particular, his decomposition of the gravitational potential into products of radial powers and spherical harmonics.

Who is a Newtonian? (Chapters 1 and 2)

Both Laplace and Fourier frame their work to be Newtonian in spirit, yet their views about the correct mathematical theory of heat differ. Within this theory, Laplace inherits Newton's *content*: the reduction of phenomena to particles interacting by forces, extended from the heavens to the molecular realm. Fourier, refusing to speculate about the nature of heat, seems to abandon such a perspective and is therefore usually considered as a proto-positivist. I argue that Fourier is no positivist, instead he extends Newton's *method* to a new mathematical context. In Chapter 2 subsection I.3 I consider this Newtonian method as a genuine third alternative to both Scholasticism and positivism and argue that Fourier's refusal of molecular explanation is not positivism but a Newtonian approach to a *mathematical* theory of a phenomenon.

Definition of 'function' (Chapters 2 and 3)

Fourier's most contested claim was that an arbitrary function can be decomposed into a trigonometric series, a claim that prompts a more systematic discussion of what a 'function' even is. Fourier's treatment of 'function' is generally considered a loose, pre-rigorous ancestor of the modern real function, later made precise by Dirichlet and others. I argue that this is wrong. Fourier's 'function' is a map between two *sequences* of values (i.e. it has a countable domain) and is closer to an infinite vector than to a map on a continuum of points. It is neither narrower nor wider than the real function: the two are not directly comparable and rest on different conceptions of the geometrical continuum. Chapter 2 subsection III.4 contains a historical exposition and in Chapter 3 I give a definition that formalizes this conception with modern machinery and study, within it, Fourier's theorem on series representation.

The Mellin transform (Chapter 4)

While the Fourier transform is the subject of countless systematic expositions, the theory of the Mellin transform $\int_{\mathbb{R}_+} f(x)x^{-s}\frac{\mathrm{d}x}{x}$ is more scattered throughout the literature. Yet, problems that are, at root, problems of *dilation* appear in many forms: ordinary differential equations of Fuchsian type (Lesch 1997), the cone calculus (Schulze 1991) and Melrose's b-calculus (Hintz 2023) all involve differential operators built from $x\partial_x$. Outside analysis, the study of algorithms asks how runtimes scale as inputs are scaled (Flajolet and Sedgewick 2009). In each of these settings, the Mellin transform plays the role that the Fourier transform plays for translation-invariant problems. Yet these works introduce the Mellin transform separately in superficially different ways. Part of the explanation is that the theory is simply more involved than that of the Fourier transform: it has more components, builds on Fourier theory, and forces choices of convention—the exact choice of the kernel, whether asymptotic behavior is studied as $x \to 0$, as $x \to \infty$, or both—on which different authors have made different choices. Chapter 4 gives a more systematic account of the Mellin transform in its own right with applications of analysis in mind: the properties of the transform on a fixed vertical line are obtained from properties of the Fourier transform by a change of variables, and its behavior as a function on the complex plane by complex analysis, the poles of the transformed function encoding asymptotic modes of the function. The subsequent application is a b-elliptic Fredholm and polyhomogeneity theorem (Theorem 4.3): solutions of b-elliptic equations on a compactified manifold admit asymptotic expansions whose modes are the boundary spectrum. Such modes are determined in part by the shape of the domain, which is the connection to the overarching theme of the thesis.

What is new and what is not

Since this thesis mixes historical exposition, philosophical argumentation, and mathematics of varying degrees of originality, I will state explicitly what I consider to be novel ideas.

- *The Society of Arcueil (Chapter 1):* exposition from primary sources with no new mathematics. My contribution is the assessment to what extent Laplace understood the spectral ideas that Fourier was to develop: he possesses the orthogonality relations, and in the urn problem even the projection formula, yet he never isolates the decomposition of functions as a subject beyond the problem at hand. Within it, I give various critical remarks, in particular the top-down/bottom-up contrast between his two decomposition arguments and the reconstruction of why his near-sphere generalization is nevertheless valid. On the Laplacian theory of heat, the discussion of Poisson's mesoscopic scale in relation to Fourier's is, as far as I am aware, not found in the literature.
- *Jean-Baptiste Joseph Fourier (Chapter 2):* the following are my own: an account of the distinction made by Fourier between the nature and mode of action of heat, with archival evidence; the argument that Fourier is Newtonian in George Smith's sense rather than a proto-positivist; and the account of Fourier's solution to the heat equation as an Aristotelian demonstration in contrast to Poisson's; the account of Fourier's function-concept as a map between two infinite sequences, irreducibly different from the real function and grounded in a different conception of the continuum; the reconstruction of the unstated premises in his proof on the existence of Fourier series decomposition.

- *A modern formalization of Fourier's 'function' (Chapter 3):* my own. The mathematical ideas exist in the literature (martingale-type conditions on a-adic refinements, cell averages), but taking it as a basic definition of 'function', that these form a Banach space, the Fourier series representation theorem, and the observation that in this context 'every function has a Fourier series' could mean three inequivalent statements, are new.
- *The Mellin Transform (Chapter 4):* no new transform theory or b-calculus results; the Fredholm and polyhomogeneity theorem is known in several frameworks. My contribution is primarily in the motivation and organization: the systematic development of the transform in its own right, the explicit statement of the trade-off between vertical decay and remainder regularity, the comparisons of the frameworks of Hintz, Schulze and Lesch, and the proofs, which are my own except for the main Theorem 4.3, whose argument follows the strategy of Hintz 2023, Theorem 3.1.
- *Moffatt eddies (Conclusion):* the result is Moffatt's (1964); its derivation here within the b-calculus framework serves as illustration of the thesis theme, not as a new result.

The thesis as a whole might serve as an illustration of how the fields of mathematics and the history of science can mutually enrich one another when studied in tandem.

Mathematical sections

Since this thesis is written for two audiences, material demanding technical mathematics has been gathered into sections marked with ⋆, so that the historical and philosophical argument should be accessible without advanced mathematical knowledge.

Acknowledgements

I thank my supervisors Viktor Blåsjö and Michał Wrochna for all the support throughout and feedback on the thesis.

Chapter 1

The Society of Arcueil

The Enlightenment saw unprecedented flourishing in all aspects of human life. In mathematics, its supreme achievement was the development of *analysis*. Originating in the calculus of Newton and Leibniz, analysis was extensively developed throughout the Enlightenment. By the turn of the 19th century, it was extensive enough to form the mathematical foundation for various major applications. Consider the following publications:

1788 (1811/5[†]) *Mécanique analytique* by Joseph-Louis Lagrange
1812 *Théorie analytique des probabilités* by Pierre-Simon Laplace
1822 (1807[‡]) *Théorie analytique de la chaleur* by Joseph Fourier

[†] Substantially expanded second edition in two volumes.

[‡] Fourier's essential ideas are already developed in his 1807 paper.

All sought to systematically apply the ideas of 'general analysis', as Fourier called it, to particular fields of knowledge. This interest in analysis was not a mere cultural fad of the time, these works introduced ideas that have been core to mathematical science ever since: Lagrange introduced action-based physics in generalized coordinates; Laplace made analysis the indispensable instrument of the probabilist; and Fourier developed the methods that earned him the title of *créateur de la physique-mathématique*.

This is a reference to the subtitle of Fourier's biography Dhombres and Robert 1998.

The first two chapters of this thesis center on Fourier's *Théorie*, where the present chapter serves as contextualization through the work of Fourier's contemporaries. To this end, we present in sections I and II two problems solved by Laplace wherein he decomposes (arbitrary) functions into so-called special functions—spherical harmonics and Hermite polynomials respectively. Fourier's primary legacy centers around the theory and application of such decompositions; contrasting his ideas with these earlier works allows us to isolate the views distinctive to Fourier. Section III considers the work of Laplace and his followers on the mathematical theory of heat, the views against which Fourier positioned his own theory.

We intentionally omit the extensively documented vibrating string controversy.

I The multipole expansion of the gravitational potential

The analytical methods did not merely recast geometric arguments in symbols, they afforded whole different methods of proof and affected the quantities which physicists considered. In particular, the *gravitational potential* was introduced for its analytical properties. 'Force' is a vector quantity: it has both a magnitude and a direction. Consequently, the analytical study of forces through a coordinate system results in three interrelated equations, for the x, y, z directions. These three equations can be reduced to a *single* equation involving a scalar quantity—the gravitational potential—so that the three equations are recovered as derivatives of this function. Laplace describes the gravitational potential generated by a spheroid as follows:

> If we designate by V the sum of all molecules of the spheroid divided by their respective distances to the attracted point, and if we call x', y', z' the coordinates of the molecule $\mathrm{d}M$ of the spheroid, and x, y, z those of the attracted point; then we will have
>
> $$V = \int \frac{\rho \,\mathrm{d}M}{\sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2}}. \tag{1}$$

Laplace seems to use spheroid here in the sense of an extended body of arbitrary shape, see † below.

Laplace 1799, Bk.3 Ch.1 §4. I added the density ρ since Laplace will do so silently later. Note that $\mathrm{d}M = \mathrm{d}x'\,\mathrm{d}y'\,\mathrm{d}z'$.

Properties about interacting (systems of) forces can be derived from their potential, motivating the study of the analytical properties of the gravitational potential. In particular, it satisfies a partial differential equation (PDE). By a simple computation, Laplace derived it:

$$\frac{\mathrm{d}^2 V}{\mathrm{d}x^2} + \frac{\mathrm{d}^2 V}{\mathrm{d}y^2} + \frac{\mathrm{d}^2 V}{\mathrm{d}z^2} = 0. \tag{2}$$

Laplace applies Eq. (2) also to interior points in §9 and §13 of the chapter we are about to discuss. In doing so he fails to notice that a source term $-4\pi\rho$ should be added on the RHS in that case.

In this section we follow Laplace's *Mécanique Céleste* Book III Chapter 2 where he determines the gravitational effects of an arbitrary spheroid on external points by developing the gravitational potential it induces into a series.

†In practice, Laplace assumes that the body is what is today called a star-shaped domain: it has a midpoint such that every ray from it intersects the boundary exactly once, so that the distance r' from the midpoint to the boundary can be expressed a function of the angular coordinates μ' and φ'.

I.1 Laplace's argument in outline

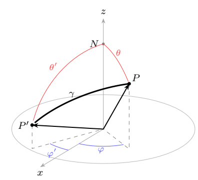
Laplace first formulates Eqs. (1) and (2) in spherical coordinates (r, θ, φ), obtaining respectively

$$V = \iiint \frac{\rho r'^2 \,\mathrm{d}r'\,\mathrm{d}\mu'\,\mathrm{d}\varphi'}{\sqrt{r'^2 - 2r'r\cos\gamma + r^2}} \tag{3}$$

and

$$0 = \frac{\mathrm{d}}{\mathrm{d}\mu}\left((1-\mu^2)\frac{\mathrm{d}V}{\mathrm{d}\mu}\right) + \frac{1}{1-\mu^2}\frac{\mathrm{d}^2 V}{\mathrm{d}\varphi^2} + r\frac{\mathrm{d}^2(rV)}{\mathrm{d}r^2}. \tag{4}$$

(Like Laplace we used $\mu = \cos\theta$, but additionally, unlike Laplace, we introduced γ: the angle between the position vector of the evaluation point and that of the source mass. γ is thus a function of the four spherical coordinates $\theta, \varphi, \theta', \varphi'$ or, what is equivalent, a function of $\mu, \varphi, \mu', \varphi'$. By the spherical law of cosines $\cos\gamma = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\varphi - \varphi')$, a basic identity obviously known to Laplace. The use of $\cos\gamma$ in Eq. (3) declutters some of Laplace's algebra without changing its essential content.)



The relationship among the various angles.

The basic technique for solving equations involving unknown functions, from Newton and Leibniz' work onwards, had been the application of power series: write the function as a power series in one of the variables, use the given (differential) equation to derive a condition on the coefficients of the power series, then identify the nature of the function by its power series. The simplicity of this description hides the mathematical difficulties of actually doing it in practice, but its universal applicability had made it a default technique for solving differential equations. Applying this general principle, it was evident to Laplace that for external points ($r > r'$) one should expand V in powers of $1/r$, i.e.

Cf. Poisson's abstract description of this method in subsection III.2.

$$V = \frac{U^{(0)}}{r} + \frac{U^{(1)}}{r^2} + \frac{U^{(2)}}{r^3} + \frac{U^{(3)}}{r^4} + \dots. \tag{5}$$

He immediately adds that it is clear from the integral expression of the potential Eq. (3) that $U^{(i)}$ is a function of μ, $\sqrt{1-\mu^2}\sin(\varphi)$ and $\sqrt{1-\mu^2}\cos(\varphi)$, it depends on the nature of the body.

$\mu = z/r$, $\sqrt{1-\mu^2}\sin(\varphi) = x/r$ and $\sqrt{1-\mu^2}\cos(\varphi) = y/r$.

The first two articles of Laplace's chapter (§8–9) reduce the triple integral defining V to triple integrals defining the individual $U^{(i)}$. The rest of the chapter then turns to simplifying the computations for the $U^{(i)}$: first for an almost spherical body (§§10–14) and then for a general body (§§15–17). Ultimately, Laplace finds that

$$U^{(i)}(\mu,\varphi) = \frac{4\pi}{(i+3)(2i+1)} \int \rho(a) \frac{\mathrm{d}\left(Z^{(i)}(a,\mu,\varphi)\right)}{\mathrm{d}a}\,\mathrm{d}a.$$

where $Z^{(i)}(a,\mu,\varphi)$ are derived from the shape of the body: Laplace slices the body into layers indexed by a, the layer a of radius $r'(a,\mu',\varphi')$ in direction μ',φ' with density $\rho(a)$. For fixed a, $Z^{(i)}$ is then the degree-i term of the decomposition of r'^{i+3} into spherical harmonics.

Laplace does not use the term 'spherical harmonics', they are defined by Eq. (11).

In retrospect we recognize in Laplace's solution a decomposition of both the potential V and the radial function r'^{i+3} into spherical harmonics. His solution exploits the fact that the spherical harmonics are eigenfunctions of the angular part of Eq. (4) (cf. Eq. (8)) to split the integral into a *universal* angular part—the same for every body, and whose integral can be computed a priori using the orthogonality of the modes—and a radial part, which is all that remains to be computed for a given shape.

I.2 Mathematical derivation

In §§8–9 Laplace derives a characterization of the $U^{(i)}$ as follows. The denominator within Eq. (3) is

$$\frac{1}{\sqrt{r^2 - 2rr'\cos\gamma + r'^2}} = \frac{1}{r} \cdot \frac{1}{\sqrt{1 - 2\frac{r'}{r}\cos\gamma + \left(\frac{r'}{r}\right)^2}}.$$

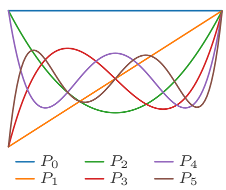
Its expansion in powers of $1/r$ is of the form

$$\frac{1}{r}\left(P_0(\cos\gamma) + P_1(\cos\gamma)\frac{r'}{r} + P_2(\cos\gamma)\frac{r'^2}{r^2} + \cdots\right). \tag{6}$$

Laplace used a generic $Q^{(i)}$ where I am using $P_i(\cos\gamma)$.

Here the P_i are *by definition* the coefficients of this expansion. They are polynomials of a single variable and are called Legendre polynomials today. Filling Eq. (6) into Eq. (3) gives

$$V = \sum_{i=0}^{\infty} \frac{1}{r^{i+1}} \iiint \rho r'^{i+2} P_i(\cos\gamma)\,\mathrm{d}r'\,\mathrm{d}\mu'\,\mathrm{d}\varphi' \tag{7}$$



The first few Legendre polynomials. $P_0(x) = 1$, $P_1(x) = x$, $P_2(x) = \frac{1}{2}(3x^2 - 1)$.

hence the $U^{(i)}$ in Eq. (5) are of the form

$$U^{(i)}(\mu,\varphi) = \iiint \rho r'^{i+2} P_i(\cos\gamma)\,\mathrm{d}r'\,\mathrm{d}\mu'\,\mathrm{d}\varphi'.$$

Note that $\gamma = \gamma(\mu,\varphi,\mu',\varphi')$ so that the integrand in Eq. (7) depends on μ' and φ'.

Filling the series decomposition Eq. (7) into Eq. (4) and equating coefficients results in distinct PDEs for each $U^{(i)}$:

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\left((1-\mu^2)\frac{\mathrm{d}U^{(i)}}{\mathrm{d}\mu}\right) + \frac{1}{1-\mu^2}\frac{\mathrm{d}^2U^{(i)}}{\mathrm{d}\varphi^2} + i(i+1)U^{(i)} = 0. \tag{8}$$

In §§10–14, Laplace considers a simplified case of 'les sphéroïdes très-peu différent d'une sphère'. In particular, he assumes the boundary to be defined by $r'(\mu',\varphi') = a(1+\alpha y(\mu',\varphi'))$ with α 'un très-petit coefficient constant, dont nous négligerons le quarré et les puissances supérieures'. Hence the *shape* is still arbitrary (since y is), but the *magnitude* of the deviation from a reference sphere is not: it is tiny compared to the radius of this sphere. This simplification—obviously motivated by the shape of actual planets—allows Laplace to avoid the integrations defining $U^{(i)}$. From the law of gravitation, he derives a boundary condition at the surface of the body. He then derives from that boundary condition, by suppressing the higher order α, that

$$4\alpha\pi a^2 y = \frac{U'^{(0)}}{a} + \frac{3U^{(1)}}{a^2} + \frac{5U^{(2)}}{a^3} + \ldots \tag{9}$$

Here $U^{(0)}$ is decomposed as $\frac{4\pi a^3}{3} + U'^{(0)}$ with Eq. (9) containing only the $\mathcal{O}(\alpha)$ part $U'^{(0)}$. Eq. (9) expresses $4\alpha\pi a^2 y$ as a series of spherical harmonics. Hence, y is itself such a series

$$y = Y^{(0)} + Y^{(1)} + Y^{(2)} + \ldots,$$

with degree-i term $Y^{(i)} = \frac{2i+1}{4\alpha\pi a^{i+3}}U^{(i)}$. As such a decomposition is unique (§12), we can equate the degree-i term of y with the degree-i term on the right-hand side of Eq. (9), which allows us to isolate $U^{(i)}$ as:

$$U^{(i)} = \frac{4\alpha\pi a^{i+3}}{2i+1} Y^{(i)}. \tag{10}$$

The $U^{(i)}$ are thus obtained from a decomposition of y, sidestepping the triple integration.

Turning now to the general case, Laplace massages the term $P_i(\cos\gamma)$ further in §15. Since $\cos\gamma = \mu\mu' + \sqrt{1-\mu^2}\sqrt{1-\mu'^2}\cos(\varphi-\varphi')$, the coordinates of the evaluation point μ,φ and the source mass μ',φ' are entangled inside the polynomial's powers. Laplace disentangles these, obtaining 'l'expression générale de $Y^{(i)}$' that satisfy Eq. (8):

$$Y^{(i)} = \sum_{n=0}^{i} (1-\mu^2)^{\frac{n}{2}} \left[\mu^{i-n} - \frac{(i-n)(i-n-1)}{2(2i-1)}\mu^{i-n-2} + \ldots\right] (a_n \sin(n\varphi) + b_n \cos(n\varphi)). \tag{11}$$

Modern readers might appreciate the notation

$$Y^{(i)}(\mu,\varphi) = \sum_{n=0}^{i} P_i^n(\mu)\,(a_n\cos(n\varphi) + b_n\sin(n\varphi)),$$

where $P_i^n(\mu)$ stand for the corresponding term in Eq. (11). The $P_i^n(\mu)$ are today known as Legendre functions of degree i and order n, up to a normalization constant which can be absorbed into a_n and b_n. In deriving this, Laplace used what today is called the *Spherical Harmonic Addition Theorem*:

$$P_i(\cos\gamma) = P_i(\mu)P_i(\mu') + 2\sum_{n=1}^{i} \frac{(i-n)!}{(i+n)!} P_i^n(\mu)P_i^n(\mu')\cos(n(\varphi - \varphi')).$$

Substituting the addition theorem into $U^{(i)} = \iiint \rho r'^{i+2} P_i(\cos\gamma)\,\mathrm{d}r'\,\mathrm{d}\mu'\,\mathrm{d}\varphi'$ and splitting $\cos(n(\varphi-\varphi')) = \cos n\varphi \cos n\varphi' + \sin n\varphi \sin n\varphi'$ separates the evaluation coordinates from the source coordinates: the factors $P_i^n(\mu)\cos n\varphi$ and $P_i^n(\mu)\sin n\varphi$ leave the integral, while the factors in μ', φ' are integrated over the body, defining the coefficients a_n and b_n.

In §16 he considers an arbitrary polynomial S of degree s in x, y and z, restricted to a sphere of radius r' after the coordinate transformation

$$x = r'\mu, \qquad y = r'\sqrt{1-\mu^2}\cos\varphi, \qquad z = r'\sqrt{1-\mu^2}\sin\varphi.$$

The result of rearranging the terms by sines and cosines of multiples of φ is a function of μ and φ with $(s+1)^2$ arbitrary coefficients. The general $Y^{(i)}$ contains $2i+1$ arbitrary constants and $\sum_{i=0}^{s}(2i+1) = (s+1)^2$ thus any S can be written as $Y^{(0)} + Y^{(1)} + \cdots + Y^{(s)}$. Laplace explicitly mentions this match of dimensions, noting that the sum of the harmonics 'renferme pareillement $(s+1)^2$ constantes indéterminées'. Afterwards he gives an explicit algorithm for computing the decomposition: subtract the most general $Y^{(s)}$ from S, choosing its $2s+1$ constants to annihilate the terms of degree s in the angular variables; the remainder is of degree $s-1$, now iterate.

The chapter climaxes in §17 where all parts developed so far come together. He recalls our goal to determine

$$U^{(i)}(\mu,\varphi) = \iiint \rho r'^{i+2} P_i(\cos\gamma)\,\mathrm{d}r'\,\mathrm{d}\mu'\,\mathrm{d}\varphi' \tag{12}$$

for an arbitrary spheroid. By the arbitrariness of the shape, the bound of integration for r' depends on μ' and φ'. Laplace first transfers this dependence to the integrand by dividing the body into nested layers—like the shells of an onion—labelled by a parameter a: the layer a is the surface $R' = R'(a, \mu', \varphi')$, with a increasing from 0 at the centre to its value at the boundary, and taken to be a surface of constant density, so that $\rho = \rho(a)$. By applying the chain rule $R'^{i+2}\,\mathrm{d}R' = \frac{1}{i+3}\frac{\mathrm{d}R'^{i+3}}{\mathrm{d}a}\,\mathrm{d}a$ to change the integration variable from the radial distance R' to the layer parameter a, we obtain:

The assumption that the body can be decomposed into layers of equal density is a slight loss of generality, Laplace uses it later to let $\rho(a)$ leave the angular integral.

$$U^{(i)}(\mu,\varphi) = \frac{1}{i+3}\iiint \rho \frac{\mathrm{d}(R'^{i+3})}{\mathrm{d}a} P_i(\cos\gamma)\,\mathrm{d}a\,\mathrm{d}\mu'\,\mathrm{d}\varphi'.$$

The arbitrariness of the body's shape now only resides in the arbitrariness of R' and not the limits of integration. Subsequently, Laplace asks us to

> Let us conceive R'^{i+3} developed in a series of this form,
>
> $$Z'^{(0)} + Z'^{(1)} + Z'^{(2)} + Z'^{(3)} + \ldots$$
>
> $Z'^{(i)}$ being, whatever i may be, a rational and integral function [i.e. a polynomial] of μ', $\sqrt{1-\mu'^2}\sin\varphi'$, and $\sqrt{1-\mu'^2}\cos\varphi'$, which satisfies the

equation in partial differences,

$$0 = \frac{\mathrm{d}}{\mathrm{d}\mu'}\left((1-\mu'^2)\frac{\mathrm{d}Z'^{(i)}}{\mathrm{d}\mu'}\right) + \frac{1}{1-\mu'^2}\frac{\mathrm{d}^2 Z'^{(i)}}{\mathrm{d}\varphi^2} + i(i+1)Z'^{(i)}.$$

Laplace 1799, Bk.3 Ch.2 §17

Laplace now uses the orthogonality of spherical harmonics (§12) to deduce

'Orthogonality' is of course not Laplace's term, but he has the integral identity which defines this relation.

$$U^{(i)}(\mu,\varphi) = \frac{1}{i+3}\iiint \rho\frac{\mathrm{d}(Z'^{(i)})}{\mathrm{d}a}P_i(\cos\gamma)\,\mathrm{d}a\,\mathrm{d}\mu'\,\mathrm{d}\varphi'.$$

The coefficient still contains an angular integral (i.e. with respect to $\mathrm{d}\mu'\,\mathrm{d}\varphi'$), which Laplace will compute. To this end he returns to the simplified case of a body 'homogène et peu différent d'une sphère', taking $\rho = 1$ and $r' = a(1+\alpha y')$. Supposing that y' (like r'^{i+3}) is developed into spherical harmonics $Y'^{(0)} + Y'^{(1)} + Y'^{(2)} + \ldots$ and suppressing terms of order α^2, he deduces

$$U^{(i)}(\mu,\varphi) = \alpha a^{i+3}\iint Y'^{(i)}(\mu',\varphi')P_i(\cos\gamma)\,\mathrm{d}\mu'\,\mathrm{d}\varphi'.$$

For clarity, I added the functional arguments in the rest of this section.

Then, recalling Eq. (10), i.e.

$$U^{(i)}(\mu,\varphi) = \frac{4\alpha\pi\, a^{i+3}}{2i+1}Y^{(i)}(\mu,\varphi)$$

he deduces 'ce résultat remarquable'

Recall that $\cos\gamma = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\varphi-\varphi')$

$$\iint Y'^{(i)}(\mu',\varphi')P_i(\cos\gamma)\,\mathrm{d}\mu'\,\mathrm{d}\varphi' = \frac{4\pi}{2i+1}Y^{(i)}(\mu,\varphi). \tag{13}$$

Today this known as the reproducing kernel property of spherical harmonics: integrating *any* degree-i harmonic against $P_i(\cos\gamma)$ scales it by $\frac{4\pi}{2i+1}$, but does not change its form. Applying Eq. (13) layer by layer evaluates the integral and collapses the coefficient to

$$U^{(i)}(\mu,\varphi) = \frac{4\pi}{(i+3)(2i+1)}\int \rho(a)\frac{\mathrm{d}\left(Z^{(i)}(a,\mu,\varphi)\right)}{\mathrm{d}a}\,\mathrm{d}a.$$

This equation is Laplace's grand prize, it reduces the triple integral of Eq. (12) to a single integral of a single term of the power series of R'^{i+3}. The rest of the chapter concerns techniques to compute it and certain simplified cases that we needn't consider. The legitimacy of generalising to arbitrary bodies an identity proved in the near-sphere case is considered in remark (ii) below.

(i) Laplace 'conceives' the arbitrary function R'^{i+3} to be decomposed in spherical harmonics. If the algorithm of §16 is the basis for such a decomposition, then it relies on the unstated premises that: (a) arbitrary means 'defined by an arbitrary power series' and (b) the decomposition of (finite) polynomials into spherical harmonics generalizes to the decomposition of (infinite) power series into spherical harmonics. The first of these seems to conflict with his view—expressed in 1779 already—that 'discontinuous functions' (in the Eulerian sense) are to be admitted in analysis; we will return to this view in Chapter 2 subsection III.4. Laplace's argument for the second contains a logical flaw. His proof for the polynomial case consists of a top-down algorithm, working from the highest order to the lowest. Consequently, the algorithm is inapplicable when the order of

This view is expressed in Laplace 1779. The chapter from the *Mécanique* originates from 1782; see Laplace 1782. I note only the coexistence of these (contradictory) views, without tracing their development in Laplace's thought.
The Eulerian notion of discontinuity is discussed in Chapter 2 subsection III.2.

the polynomial becomes infinite. The only ground he can stand on is the floating abstraction that since both involve infinitely many arbitrary constants, the decomposition should be possible. Analytical definitions and proofs generally involve some kind of successive procedure, but valid successive procedures are bottom-up, e.g. the derivative and the (Riemann) integral are defined through a successive refinement of finite differences and sums respectively. The bottom-up requirement is implicit in the definition of a limit: for any $\varepsilon > 0$ one can specify a number of steps N which suffices. A bottom-up argument might still fail, e.g. when a quantity does not converge as n increases, but an argument like Laplace's can be rejected a priori by a critical reader.

A valid bottom-up procedure does exist: the degree-i coefficient is obtained by orthogonal projection (as we would say), like for trigonometric series. If we allow, as Laplace would certainly have, interchanging infinite sums and integration then this coefficient formula is implicit in Eq. (13) together with the orthogonality proven in §12. Laplace has the mathematical formulae, yet he does not invoke them as a method for decomposing arbitrary functions. In section II, we will see Laplace use the orthogonal projection formula for a different problem.

(ii) Laplace makes *another* generalization which might seem unwarranted. Having used the particular case of a solid 'un peu différent d'une sphère' to derive Eq. (13), he proclaims it generally. However, this turns out to be a valid argument. The smallness assumption pertains to α, and does not place restrictions on the form of y'. For the fixed near-sphere body $r' = a(1+\alpha y')$ the coefficient $U^{(i)}$ is a function of α that Laplace computed two ways: through the body-integral and through a surface condition. We thus have:

Of course y' has to be bounded for $\alpha y'$ to be small, something tacitly assumed by Laplace.

$$U^{(i)} = \alpha A + \mathcal{O}(\alpha^2), \qquad U^{(i)} = \alpha B + \mathcal{O}(\alpha^2),$$

with $A = a^{i+3} \iint Y'^{(i)} P_i(\cos\gamma)\,\mathrm{d}\mu'\,\mathrm{d}\varphi'$ and $B = \frac{4\pi a^{i+3}}{2i+1} Y^{(i)}$. Two expansions of one function share their first-order coefficient, so $A = B$. Laplace would later offer such comparisons as the basis of the rigor of infinitesimal calculus, writing in 1810:

> The quantities that are neglected in these transitions from the finite to the infinitely small seem to deprive infinitesimal calculus of the rigor of geometrical results; but, to restore it, it is sufficient to regard the quantities retained in the expansion of an equation in finite differences and of its integral, with respect to the powers of the indeterminate difference, as all having for a factor the smallest power whose coefficients are compared with one another. Since this comparison is rigorous, differential calculus, which is clearly nothing other than this very comparison, possesses all the rigor of the other algebraic operations.

Laplace 1810a, pp.358–9

I.3 General remarks

Many things can be said about Laplace's impressive but difficult chapter, I will restrict myself to the themes of this thesis.

(1) Laplace's decomposition of the potential V into terms of the form $U^{(i)}/r^{i+1}$ is derived from the integral expression of the potential (Eq. (3)). Laplace does not posit the terms $U^{(i)}/r^{i+1}$ as separated solutions of a PDE

(Eq. (4)) that are superposed to define a more general solution. This makes sense in this particular problem, since Laplace's starting point was the integral expression for V and not the PDE. However, in the debates about proper method for solving the heat equation—which is first and foremost a PDE—this will become a focal point of disagreement. Can, as Fourier did, the series solutions be derived by composing particular 'simple modes' that solve the problem? Or, as Poisson argued against Fourier, does the series solution have to be derived by developing the integral of the PDE into a series? In that case, a general integral has to be derived first.

We will discuss Poisson's views in subsection III.2.

Additionally, what is the purpose of the series decomposition? Is the decomposition purely a computational device, or is it more than that? The mere application of a formal mathematical technique to decompose a quantity does not endow the constituent parts with physical significance; the individual terms of a decomposition—like the digits of a decimal number—might serve a purely computational goal. Fourier, however, believes that the terms of his series are more than the digits of a decimal expansion, writing that 'the natural effect whose laws we seek, is really decomposed into distinct parts, corresponding to the different terms of the series' and that the 'essential character' of his solution is the 'representation of phenomena'. Fourier stresses that his solutions 'serve to determine with facility all their results numerically', however, the 'simple modes' that compose his solutions are *more* than that, since 'without determining the values of the coefficients, we can always acquire an exact knowledge of the problem, and of the natural course of the phenomenon which is its object; the chief consideration is that of *simple modes*'.

Théorie §428; these passages will be discussed in Chapter 2 subsection II.4.

Fourier uses the term 'mouvements simples', throughout the thesis I have translated this as 'simple mode'. He uses 'mode' instead of 'mouvement' once in §246.

Laplace is clearly aware that the individual $U^{(i)}$ have relations to physical properties: he knows that $U^{(0)}$ corresponds to the total mass (§11), $U^{(1)}$ to the offset of the center of mass (§12) and that odd terms vanish for an equatorially symmetric body (§17). Moreover, he shows that, for a sphere deformed by a single harmonic $y = Y^{(i)}$, the potential produced by the deformation is that same harmonic rescaled, and that this correspondence holds only for such deformations (§11). He argues that the decomposition of y into spherical harmonics is 'n'est donc point arbitraire, mais elle dérive du développement en série des attractions des sphéroïdes' (§11) and derives that y 'ne peut se développer que d'une seule manière' from the orthogonality of the harmonics: $\iint Y^{(i)} Z^{(i')} \,\mathrm{d}\mu\, \mathrm{d}\varphi = 0$ for $i \neq i'$ (§12). However, he does not pause to consider the spherical harmonics and the decomposition of functions into them any more than is necessary to solve the particular problem at hand: the computation of the integral within this particular problem concerning gravity, leading Isaac Todhunter to complain that the chapter is not 'well arranged' since 'the pure analysis and the physical application of it are not kept sufficiently distinct'. When developing various different quantities into spherical harmonics, their terms are expressed by a symbol for the term of the power series of the quantity they were derived from—$U^{(i)}$ for V, $Y^{(i)}$ for y, $Z^{(i)}$ for R'^{i+3} and various others—rather than explicitly recognizing that all these are of the same form and using expressions

See Fig. 1 for the meaning of the $U^{(i)}$.

Todhunter 1873, §1077. Isaac Todhunter (1820–1884) is a Victorian mathematician known for his exhaustive historical compilations of primary sources.

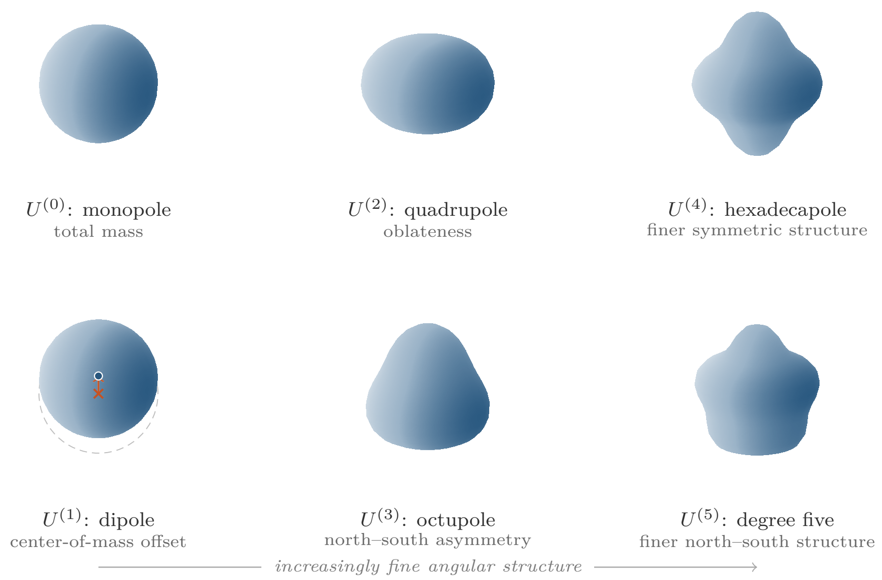


Figure 1. Representative forms of the first six multipole degrees. The surfaces are radial perturbations proportional to $P_i(\cos\theta)$, with the dipole representing a displacement to first order. Increasing degree introduces progressively finer angular structure. Only one representative is shown for each degree; a general $U^{(i)}$ is a linear combination of all spherical harmonics of degree i.

like $u_i\phi_i$, $y_i\phi_i$, $z_i\phi_i$ with u_i, y_i, z_i numbers and ϕ_i the common spherical harmonic. Fourier will use a single symbol for the function he develops other functions into (see Chapter 2 subsection III.3) and his *Théorie* has multiple sections specifically dedicated to series decompositions (see Chapter 2 section III).

It is thus unclear whether Laplace considers it an accidental feature of his computational goal that the $U^{(i)}$ have a physical significance or whether *the purpose* of the decomposition is the description of distinct physical effects in terms of its 'elementary modes', with the computational power being a consequence.

This is again Fourier's term, *Théorie* §171. I again translated 'mouvement' as 'mode'.

(2) At root, Laplace derives the reproducing kernel identity Eq. (13) from a boundary condition, which is itself derived from the inverse square law of gravitation. This means that Laplace's argument for the decomposition of the potential function V into a power series in $1/r$ with spherical harmonics as coefficients *depends* on the physical nature of gravity. Laplace seems to be aware of this, starting his derivation of the boundary condition with a more general integral over *some* power $n+1$ of the distance f of the source point and the evaluation point, then stating 'dans le cas de la nature, où $n=-2$, elle devient $\int \frac{\mathrm{d}M}{f}$'. Laplace concieved of different physical laws governed by different powers, what Laplace did not conceive is that the measure of *distance itself*—not merely the exponent in the force law—could vary. That is the premise of general relativity, where the metric is dynamical, and it removes the foundation for expansions

into integer powers: on a general manifold, one does expand into powers of the radius, however these powers need not be integers (they can be irrational or even complex). The tool that identifies these powers is the Mellin transform, taken up in Chapter 4.

II The scaling limit of the Bernoulli–Laplace urn

In this section we will consider a problem from Laplace's *Théorie analytique des probabilités*. We will again see Laplace develop an 'arbitrary' function into a series of special functions, in this case the Hermite polynomials. Like in the problem of the previous section, these decompositions emerge within the solution of a particular problem. Laplace again derives the series from an integral expression, but since the problem takes off from a PDE in this case, that integral has to be derived first. New in this problem is Laplace's derivation and use of orthogonal projection formula to determine the coefficients that decompose the initial state into the Hermite polynomials. What is again absent is an independent consideration of the decomposition of (arbitrary) functions into Hermite polynomials beyond what is necessary to compute the solution to the problem at hand.

Laplace of course does not call them Hermite polynomials.

II.1 Laplace's argument in outline

Laplace discusses the following problem in Book II Chapter 3: Two urns A and B each contain n balls and of the total $2n$ balls half is black and half white. At each step, interchange two balls (each ball in the urn has an equal chance of being drawn). Sought is

$$\begin{aligned} z_{x,r} &= \mathbb{P}\{\text{There are } x \text{ white balls in } A \text{ after } r \text{ swaps}\} \\ &= \frac{\#\text{sequences of } r \text{ swaps yielding } x \text{ white in } A}{\#\text{sequences of } r \text{ swaps}}. \end{aligned}$$

Mathematically, this is a purely combinatorial problem: given an initial distribution of the balls, count the number of sequences of swaps yielding x white balls in A after r steps and divide it by the total amount of sequences of r swaps.

Laplace first transforms the problem into one he can actually solve: by taking n 'a very great number' (so that we can ignore terms of order $1/n^2$) and under a change of variables $x = \frac{n+\mu\sqrt{n}}{2}$, $r = nr'$ and $z_{x,r} = u(\mu, r')$—obviously motivated by the central limit theorem—he derives a PDE that governs the evolution of the probabilities from some initial probability distribution:

$$\frac{\mathrm{d}u}{\mathrm{d}r'} = 2u + 2\mu\frac{\mathrm{d}u}{\mathrm{d}\mu} + \frac{\mathrm{d}^2u}{\mathrm{d}\mu^2}. \tag{14}$$

He then integrates this PDE and subsequently develops u into a series, which we can in retrospect identify as the Hermite polynomials H_k. He then obtains a formula for the coefficients of this series in terms of an

The H_k are defined in Eq. (24).

initial state $X(\mu)$. Using our notation instead of Laplace's, he obtains:

$$u(\mu,r') = \frac{2e^{-\mu^2}}{\sqrt{n\pi}} \sum_{k=0}^{\infty} \frac{c_k}{k!} e^{-2kr'} H_k(\mu)$$

$$c_k = \frac{1}{\int H_k(\mu)H_k(\mu)e^{-\mu^2}\,\mathrm{d}\mu} \int X(\mu)H_k(\mu)e^{-\mu^2}\,\mathrm{d}\mu.$$

★ II.2 Mathematical derivation

At each step, there are n^2 possible swaps so that n^{2r} sequences of swaps can arise from r steps. Laplace subsequently relates the number of ways $\{x$ white balls in A after $r+1$ steps$\}$ can arise to the distribution after r steps, obtaining a partial finite difference equation for $z_{x,r}$. This state can come about from four possible events:

- With x whites in A, swap two whites: $x(n-x)n^{2r}z_{x,r}$ possible ways.
- With x whites in A, swap two blacks: $(n-x)xn^{2r}z_{x,r}$ possible ways.
- With $x-1$ whites in A, swap a black of A for a white of B: $(n-x+1)^2n^{2r}z_{x-1,r}$ possible ways.
- With $x+1$ whites in A, swap a white of A for a black of B: $(x+1)^2n^{2r}z_{x+1,r}$ possible ways.

Their sum equals the amount of ways $\{x$ white balls in A after $r+1$ steps$\}$ can arise, i.e. $n^{2r+2}z_{x,r+1}$. Thus

$$z_{x,r+1} = \left(\frac{x+1}{n}\right)^2 z_{x+1,r} + \frac{2x(n-x)}{n^2} z_{x,r} + \left(1-\frac{x-1}{n}\right)^2 z_{x-1,r}. \tag{15}$$

Laplace observes that this difference equation determines $z_{x,r}$ for any r when the initial $z_{x,0}$ is given. However, this is a difference equation and the computations to solve it explicitly become infeasible as r grows. To transform the (difficult) finite difference equation into an (easier) differential equation, Laplace considers n 'a very great number' so that we have 'very nearly':

$$\begin{aligned} z_{x+1,r} &= z_{x,r} + \frac{\mathrm{d}z_{x,r}}{\mathrm{d}x} + \frac{1}{2}\frac{\mathrm{d}^2 z_{x,r}}{\mathrm{d}x^2} \\ z_{x-1,r} &= z_{x,r} - \frac{\mathrm{d}z_{x,r}}{\mathrm{d}x} + \frac{1}{2}\frac{\mathrm{d}^2 z_{x,r}}{\mathrm{d}x^2} \\ z_{x,r+1} &= z_{x,r} + \frac{\mathrm{d}z_{x,r}}{\mathrm{d}r}. \end{aligned} \tag{16}$$

He then introduces the change of variables

$$x = \frac{n+\mu\sqrt{n}}{2}, \qquad r = nr', \qquad z_{x,r} = u(\mu,r'), \tag{17}$$

after which he immediately writes down the PDE:

$$\frac{\mathrm{d}u}{\mathrm{d}r'} = 2u + 2\mu\frac{\mathrm{d}u}{\mathrm{d}\mu} + \frac{\mathrm{d}^2u}{\mathrm{d}\mu^2}, \tag{18}$$

See remark (ii) below for the derivation of Eq. (18).

while noting that the $1/n^2$ terms are suppressed.

To integrate the PDE Eq. (18), Laplace starts by introducing the following ansatz

$$u(\mu,r') = \int \phi(t,r')e^{-\mu t}\,\mathrm{d}t. \tag{19}$$

Substituting this back into the PDE, differentiating with respect to r' under the integral sign and applying integration by parts to the μ derivatives results in

$$\int e^{-\mu t}\frac{\mathrm{d}u}{\mathrm{d}r'}\,\mathrm{d}t = 2e^{-\mu t}t\phi + \int e^{-\mu t}\left(t^2\phi - 2t\frac{\mathrm{d}\phi}{\mathrm{d}t}\right)\mathrm{d}t.$$

Setting the resulting expression to zero gives

$$\frac{\mathrm{d}\phi}{\mathrm{d}r'} = t^2\phi - 2t\frac{\mathrm{d}\phi}{\mathrm{d}t}$$

whenever $t\phi(t,r)e^{-\mu t} = 0$ holds at the integration bounds of Eq. (19). Laplace immediately states, without elaboration, its solution

$$\phi(t,r) = e^{\frac{1}{4}t^2}\psi\left(\frac{t}{e^{2r'}}\right) \tag{20}$$

Eq. (20) can be obtained by the so-called method of characteristics.

'$\psi\left(\frac{t}{e^{2r'}}\right)$ being an arbitrary function of $\frac{t}{e^{2r'}}$'. Therefore

$$u(\mu,r') = \int e^{-\mu t+\frac{1}{4}t^2}\psi\left(\frac{t}{e^{2r'}}\right)\mathrm{d}t. \tag{21}$$

Laplace now substitutes the complex $t = 2\mu + 2s\mathrm{i}$, stating that we integrate from $s = -\infty$ to $s = \infty$ to obtain Laplace's general solution:

$$u(\mu,r') = e^{-\mu^2}\int e^{-s^2}\Psi\left(\frac{s+\mu\mathrm{i}}{e^{2r'}}\right)\mathrm{d}s \tag{22}$$

with $\Psi(z) = 2\mathrm{i}\psi(z\mathrm{i})$.

The second part of the solving the problem consists of relating the initial state of the urn to $\Psi(\frac{s+\mu\mathrm{i}}{e^{r'}})$. Laplace starts, as usual, by taking a power series expansion of Ψ, its n-th term being

$$c_n\frac{e^{-\mu^2}}{e^{2nr'}}\int e^{-s^2}(s-\mu\mathrm{i})^n\,\mathrm{d}s. \tag{23}$$

Laplace works out the integrations, including calculating the normalization factor needed so that $\int u\,\mathrm{d}x = \sqrt{n}/2\int u\,\mathrm{d}\mu = 1$. Here he justifies taking the integral from $\mu = -\infty$ to $\mu = \infty$ noting that 'this difference is insensible when n is a great number'. This leads to 'the general expression of u'

$$u(\mu,r') = \frac{2e^{-\mu^2}}{\sqrt{n\pi}}\left\{\begin{aligned}&c_0 + c_2(1-2\mu^2)e^{-4r'} + c_4\left(1-4\mu^2+\frac{4}{3}\mu^4\right)e^{-8r'} + \cdots\\ &+c_1\mu e^{-2r'} + c_3\mu\left(1-\frac{2}{3}\mu^2\right)e^{-6r'} + c_5\mu\left(1-\frac{4}{3}\mu^2+\frac{4}{15}\mu^4\right)e^{-10r'} + \cdots\end{aligned}\right\}.$$

Contemporary readers will appreciate the abbreviated form

$$u(\mu,r') = \frac{2e^{-\mu^2}}{\sqrt{n\pi}}\sum_{k=0}^{\infty}\frac{c_k}{k!}e^{-2kr'}H_k(\mu)$$

in terms of the Hermite polynomials

$$H_k(\mu) := k!\sum_{m=0}^{\lfloor k/2\rfloor}\frac{(-1)^m}{m!(k-2m)!}(2\mu)^{k-2m}. \tag{24}$$

He then proves that

$$\int_{-\infty}^{+\infty}H_k(\mu)H_j(\mu)e^{-\mu^2}\,\mathrm{d}\mu = 0 \qquad \text{whenever } k \neq j. \tag{25}$$

Supposing $u(\mu,0)=X(\mu)$, he completes his solution by deducing

$$c_k = \frac{1}{\int H_k(\mu)H_k(\mu)e^{-\mu^2}\,\mathrm{d}\mu} \int X(\mu)H_k(\mu)e^{-\mu^2}\,\mathrm{d}\mu. \tag{26}$$

He subsequently applies his general solution to particular initial conditions. First, he considers the initial condition $X = \frac{2k}{\sqrt{n\pi}}e^{-k^2\mu^2}$. Note that this the case that the initial white count in A is normally distributed about $n/2$ with variance $n/(8k^2)$. He shows that it retains its shape, stating that it 'becomes in its limit, when r' is infinity, $u = \frac{2}{\sqrt{n\pi}}e^{-\mu^2}$', he also treats displaced normal distributions, showing that they drift back to normal distributions centered on μ. I won't discuss these consequences in more detail and instead make various remarks on Laplace's derivation and my exposition thereof.

Laplace does not use the term 'normal distribution'.

(i) I have stayed true to Laplace's original presentation, only changing some symbols (like his use of c for the number e). In particular, I kept the same notation for finite and infinitesimal differences, which is conceptually significant; see remark (1) below. I made two more substantial changes: Laplace judiciously distinguishes the odd and even polynomials that make up the series. He treats them separately in the proof and uses different constants for their coefficients, i.e. for each n there is an odd and even term, I did not make this distinction. Also, after introducing H_k for Hermite polynomials I used them to express the subsequent decomposition of u, Laplace instead retains the U_i that stand for the coefficient of u developed into a power series.

(ii) Eq. (18) is derived as follows. Observing that $\mathrm{d}x = \frac{\sqrt{n}}{2}\mathrm{d}\mu$ and $\mathrm{d}r = n\,\mathrm{d}r'$, Eq. (16) becomes

$$\begin{aligned} z_{x+1,r} &= u + \frac{2}{\sqrt{n}}\frac{\mathrm{d}u}{\mathrm{d}\mu} + \frac{2}{n}\frac{\mathrm{d}^2u}{\mathrm{d}\mu^2} \\ z_{x-1,r} &= u - \frac{2}{\sqrt{n}}\frac{\mathrm{d}u}{\mathrm{d}\mu} + \frac{2}{n}\frac{\mathrm{d}^2u}{\mathrm{d}\mu^2} \\ z_{x,r+1} &= u + \frac{1}{n}\frac{\mathrm{d}u}{\mathrm{d}r'}. \end{aligned} \tag{27}$$

Rewriting the partial difference equation Eq. (15) as

$$z_{x,r+1} = \left(p+\frac{1}{n}\right)^2 z_{x+1,r} + 2pqz_{x,r} + \left(q+\frac{1}{n}\right)^2 z_{x-1,r}$$

$p = x/n$
$q = 1-x/n$

and filling in Eq. (27) gives

$$
\begin{aligned}
u+\frac{1}{n}\frac{\mathrm{d}u}{\mathrm{d}r'} = {} & \left(p^2+2\frac{p}{n}+\frac{1}{n^2}\right)\left(u+\frac{2}{\sqrt{n}}\frac{\mathrm{d}u}{\mathrm{d}\mu}+\frac{2}{n}\frac{\mathrm{d}^2u}{\mathrm{d}\mu^2}\right)\\
& +2pqu\\
& +\left(q^2+2\frac{q}{n}+\frac{1}{n^2}\right)\left(u-\frac{2}{\sqrt{n}}\frac{\mathrm{d}u}{\mathrm{d}\mu}+\frac{2}{n}\frac{\mathrm{d}^2u}{\mathrm{d}\mu^2}\right)\\
= {} & (p^2+2pq+q^2)\,u+(p^2-q^2)\,\frac{2}{\sqrt{n}}\frac{\mathrm{d}u}{\mathrm{d}\mu}+(p^2+q^2)\,\frac{2}{n}\frac{\mathrm{d}^2u}{\mathrm{d}\mu^2}\\
& +\frac{2(p+q)}{n}\left(u+\frac{2}{n}\frac{\mathrm{d}^2u}{\mathrm{d}\mu^2}\right)+\frac{2(p-q)}{n}\frac{2}{\sqrt{n}}\frac{\mathrm{d}u}{\mathrm{d}\mu}+\mathcal{O}(n^{-2}) \qquad \begin{matrix} p+q=1 \\ p-q=\mu/\sqrt{n} \end{matrix}\\
= {} & u+\frac{1}{n}\left(2u+2\mu\frac{\mathrm{d}u}{\mathrm{d}\mu}+\frac{\mathrm{d}^2u}{\mathrm{d}\mu^2}\right)+\mathcal{O}(n^{-2}) \qquad (28)
\end{aligned}
$$

from which Eq. (18) follows by considering the terms up to $\mathcal{O}(1/n)$.

(iii) The change of variables of Eq. (17) is of course motivated by the central limit theorem, which also justifies the approximations: Eq. (27) approximated u (i.e. z) by its by the first terms of its power series, suppressing (among others) terms involving $\frac{1}{n^2}\frac{\mathrm{d}^2u}{\mathrm{d}r'^2}$ and $\frac{1}{n^{3/2}}\frac{\mathrm{d}^3u}{\mathrm{d}\mu^3}$; these approximations break down if μ is of order $\sqrt{n}$, but by the CLT we expect the relevant cases to correspond to $\mu\approx 0$. Looking at the solution Eq. (22), we can see that $\frac{\mathrm{d}^nu}{\mathrm{d}r'^n}\propto\mu^{2n}u$ and $\frac{\mathrm{d}^nu}{\mathrm{d}\mu^n}\propto\mu^nu$ so that for $\mu\approx\pm\sqrt{n}$ the supressed terms become as large as the terms retained in the PDE. (The suppression of $\mathcal{O}(1/n^2)$ in Eq. (28) only breaks down for μ near n.) $\mu\approx\pm\sqrt{n}$ means $x\approx 0$ or n, which, in our original problem, corresponds to the probability that urn A contains (almost) only white or black balls, a very low probability event. The PDE cannot be used to estimate (very low) probability events, a price we pay to gain the machinery of analysis.

There is another way in which low probability events aren't captured by the PDE. As Laplace notices, the original problem has two boundary conditions: $x<0$ and $n<x$ (i.e. $\mu>\pm\sqrt{n}$) have zero probability, i.e. u should be subject to a BC at $\mu=\pm\sqrt{n}$ to conserve probability mass. Laplace doesn't impose such boundary conditions on the PDE and will solve the problem on the whole number line. Adding boundary conditions would greatly complicate solving the PDE. Since u is not representative there anyways, this reinforces the same limitations just discussed. Laplace thus gets rid of the boundary conditions implied by the actual problem.

(iv) Laplace's use of the integral deserves commentary. His integral is not inseparably linked to a domain of integration, as it is today; instead, the integral is first and foremost indefinite and its evaluation at definite boundary points is something you impose on it. It is therefore first and foremost the antiderivative operator. Today the definite integral is basic and that the indefinite integral is also the antiderivative is its fundamental property established by the Fundamental Theorem of Calculus. When he thus defines Eq. (19), next to the fact that ϕ is unspecified, there are no associated bounds; the issue is reversed: we should find ϕ and boundary points from the problem to make this well-defined. Eq. (19) is simply an ansatz for the solution. Eq. (20) is a discovery and from Eq. (21) we discover that we should integrate from $2\mu - \infty\mathrm{i}$ to $2\mu + \infty\mathrm{i}$ (and that we cannot integrate from $t = -\infty$ to $t = \infty$). The choice for the bounds was dictated by the boundary condition $t\phi(t, r)e^{-\mu t} = 0$, this condition has a necessary relation to Eq. (19): the expression is t times the integrand, forcing the boundary term to vanish as $|t| \to \infty$ guarantees that the integrand decays strictly faster than $\mathcal{O}(1/t)$. Consequently, the vanishing boundary terms enforce the convergence of the improper integral.

The practice of setting up an integral equation and subsequently looking for bounds where it is well-defined is far from dead today. The theory of the Mellin transform in Chapter 4 is essentially about identifying such bounds for Eq. (19). As a preview to Chapter 4, note that integrating over an arbitrary imaginary line $a + \omega\mathrm{i}$ turns Eq. (19) into $\int e^{-a}\phi(a + \omega\mathrm{i}, r')e^{-\mathrm{i}\mu\omega}\,\mathrm{d}\omega$ which is the inverse Fourier transform of $\omega \mapsto e^{-a}\phi(a + \omega\mathrm{i}, r')$ (up to a normalization constant). Now, the choice of $t = 2\mu + 2s\mathrm{i}$ was dictated by the condition that (t times) the integrand of Eq. (21) decayed at the bounds of integration, which was resolved by completing the square to exploit the decay of e^{-s^2}. From the algebraic computations it is obvious that the 2 in the real part of t comes directly from the two 2's in the original PDE Eq. (19), i.e. it would be k if the respective terms in the PDE had k as a coefficient. This choice is therefore not arbitrary and has a necessary relation with the PDE (and thus the underlying probabilistic problem).

Laplace ignores whether $\psi(1/e^{2r'})$ could counter the decay of e^{-s^2}. Note that for large r', $\psi(1/e^{2r'})$ is essentially constant.

(v) Laplace's solution involved expressing an arbitrary function in terms of Hermite polynomials. Taking the analyticity of an 'arbitrary functions' for granted, this decomposition is essentially different from the decomposition into spherical harmonics of section I: it is bottom-up and not top-down. The Hermite polynomials H_k are constructed successively in terms of the first k terms of the powers of μ and the coefficients are derived as orthogonal projections in the coefficient formula Eq. (26). Laplace does not go into detail on what kind of functions can be decomposed into these H_k. Early in the problem he claims the initial state to be 'arbitrary', but he never explicitly claims that an arbitrary function can be decomposed into the H_k. If he had, then he might have been faced with similar objections that Fourier faced in his decomposition of arbitrary functions into trigonometric series. For polynomial initial states the validity of decomposition would probably have been self-evident to anyone at the time: the Hermite polynomials are themselves polynomials, hence, an argument involving the matching of degrees applies. However, this still leaves unresolved the exact generality of such decompositions. In retrospect, we note that the Hermite polynomials are an orthogonal basis for the Hilbert space $L^2(\mathbb{R}, e^{-x^2}\,\mathrm{d}x)$.

II.3 General remarks

(1) Laplace has no separate notation for finite or infinitesimal differences, using d for both. Since this is conceptually significant I copied this practice. Throughout the derivation of the solution he works with n being a (finite) 'great number' and he never takes n to be infinite. However, he does refer to the PDE as an equation in 'partial differentials',

When *applying* the obtained formulae he does take $r' = rn$ to be infinite. However, this refers to letting r—which plays to role of time—to be infinite rather than n. Otherwise all formulae he subsequently states, like $\frac{2}{\sqrt{n\pi}}e^{-\mu^2}$, become zero.

which would suggest that n is infinite. In any case, Laplace is working with two scales in this problem: order n and order 1, and the existence of these two separate scales justifies his approximations. Mathematically this is of course equivalent to a more traditional geometric situation involving scales of order 1 and $\mathrm{d}x = 1/n$, e.g. a finite area and its tiny parts. However, while a continuous line might be infinitely divisible, the number of balls in the urn is discrete. Laplace does not hesitate to apply the tools of infinitesimal analysis to processes that are at root discrete. Let us recall the justification expressed by Laplace in remark (ii) of the previous section: the infinitesimal calculus consists in considering a finite *indeterminate* difference, developing an equation in finite differences (and its integral) in powers of that difference, and comparing the coefficients of like order; subsequently, 'l'omission des infiniment petits d'un ordre supérieur à celui que l'on conserve' then 'simplifient extrêmement les calculs'. In the context of Fourier's views of 'arbitrary functions', we will discuss how Laplace urges us to use the analysis of finite differences to settle what kind of 'arbitrary functions' could enter into infinitesimal analysis.

Laplace 1810a, p.359

See Chapter 2 subsection III.4

(2) We can recognize, in retrospect, ideas of statistical mechanics in Laplace's problem: *reversible* microscopic dynamics—the swapping of balls—gives rise to a macroscopically irreversible process. A close relative to Laplace's urn model is the Ehrenfest urn model, introduced in 1907 to defend Boltzmann's statistical account of the second law of thermodynamics against two specific objections: that the underlying dynamics is time-reversible, and that the system must eventually return arbitrarily near any earlier state. For heat, the Laplacian programme would consist of deducing macroscopic behaviour from the action of molecules; today one would do this by the very type of probabilistic limit Laplace carries out here. Laplace possessed mathematical ideas that we would classify under statistical mechanics today without (as far as I am aware) applying these to model physical processes. As we will see in subsection III.1, Poisson *does* introduce a mesoscopic scale with the expressed purpose of bridging the gap between the microscopic irregularity of molecular heat exchange and the regular behavior of macroscopic heat flow.

P. Ehrenfest and T. Ehrenfest 1907; the objections are Loschmidt's reversibility paradox and Zermelo's recurrence objection.

(3) In Chapter 4 subsection I.2 we return to this problem to gain a deeper perspective—using modern machinery—on what makes this particular problem difficult (compared to, say, the heat equation on the same domain.)

III The Laplacians on the mathematics of heat

The subject of this section and the next chapter is the mathematical theory of the movement of heat, as developed by two opposing schools: Laplace and his followers (most notably Poisson), and Fourier. I will refer to the former as the *Laplacians*. We consider the following problem: given initial and boundary temperatures, to determine mathematically the temperature inside a solid of a given material. Our discussion thus centers on the phenomenon of heat *conduction* which is described by (what is today known as) the heat equation. Other aspects of heat—its nature, the movement of heat in liquids and gasses, radiant heat, effects of heat like expansion and elasticity, the experimental determination of the relevant properties like specific heat for particular materials, and the application of the theory to the system of the world, like temperatures on earth, etc.—will enter insofar as these authors themselves take them to be necessary to establish and solve the heat equation. Doing so isolates the question at the root of the controversy: what physical and mathematical facts form a rigorous derivation and solution of the heat equation?

We present in this section the Laplacian theory and reserve Fourier's ideas for Chapter 2. This is possible, even though these camps were in dialogue and reacting to each other, because they present two essentially distinct approaches to the mathematical theory of heat. Interestingly, their disagreements were primarily methodological—how to derive and solve the heat equation—with remarkably little disagreement about content—the heat equation itself and different forms of its solution, they understood how these different forms could be derived from each other. Poisson's mature treatise *Théorie Mathématique de la Chaleur*, published in 1835, does take inspiration from Fourier's 1822 *Théorie*, but without repudiating its Laplacian soul.

III.1 Deriving the heat equation: Laplace's hypothesis

The mathematical theory of heat with which we are concerned has its origins in the work of Newton. The first experiments about the cooling of small discrete bodies were reported by Newton in an article *Scala Graduum Caloris*, published anonymously in the year 1701. From the observation that 'the temperatures decrease in a geometric progression for arithmetical progressions of time', Newton drew the obvious conclusion that small bodies lose heat proportional to their temperature. In analytic symbols this of course comes down to inferring the ODE $\mathrm{d}u/\mathrm{d}t = -au$ from observing $u(t) = e^{-at}$, where a depends on the nature of the substance and the environment. This basic principle about the rate at which (small) bodies cool is known as *Newton's law of cooling* and it is the physical basis for the theories of the Laplacians and Fourier.

See Dhombres and Robert 1998, p.445.

The second important experiment involved heating a long but thin metal bar at one end and measuring the temperatures at various points when the bar reached equilibrium. The experimental data was reported on by Amontons in 1703, but Lambert was the first to deduce the right temperature distribution from the data. In such a state, the temperatures exhibited a geometric progression, now for arithmetical progressions of distance to the source. That the same exponential curve described heat flow over time in small discrete bodies and the interior heat flow of a continuous bar motivated the idea that something like Newton's law applies to the communication of heat in the interior of solids too. The first problem of mathematizing heat treated by Laplace and Fourier consists in giving a precise account of this relationship—to explain the temperature progression observed in the bar by means of Newton's law of cooling.

Amontons measured the temperatures by placing different substances on the bar and noting the distance from the heat source where they changed state. For example, he recorded that butter melted at 42 inches, drops of water boiled at 22 inches, etc. The geometric progression was not immediately obvious until Lambert later mapped this data to a precise temperature scale. See Dhombres and Robert 1998, p.447–451 for an account.

Biot took up the problem in 1804, attempting to derive the heat equation and subsequently solve it. He seems to have understood what the right equation had to be, but his derivation of it ran into an 'analytical difficulty'. In 1808 Laplace read a note to the *Institut* in which he attempted to identify why Biot's analytical derivation failed and proposing a solution. The problem Laplace considers is slightly different from Biot's but it would involve same 'analytical difficulty'. Laplace begins with the

The note is reprinted in Laplace 1810b

Laplace considers a bar isolated at its boundaries, rather than a bar where heat moves both in the interior and at the boundary.

> principle, given by Newton, namely, that the heat communicated by one body to another contiguous to it is proportional to the difference in their temperatures. Thus an infinitely thin plate of a body communicates, in a given very short time, to the one that follows it, a quantity of heat proportional to the conductivity of the body for heat and to the excess of its temperature over that of the next plate: but it receives at the same time from the plate that precedes it a quantity of heat proportional to the excess of the temperature of this plate over its own, and it is the difference of these heats received and communicated in an infinitely small which forms the differential of its heat.

Laplace 1810b, p.290

In other words, for three consecutive slices with respective temperatures $u_{,}, u, u'$ the heat communicated to u equals

$$K(u_{,} - u) - K(u - u') \tag{29}$$

during dt. He continues to write that

> here a difficulty arises which has not yet been solved. The quantities of heat received and communicated in a moment may be infinitely small of the same order as the excess of temperature of one plate over that of the plate which follows it. The difference between the heats received and communicated is therefore an infinitely small of the second order, the accumulation of which in a finite time could not raise the temperature of the plate by a finite amount.

Eq. (29) is absurd: $u_{,} - u$ and $u - u'$ are first-order differentials, their difference is a second order differential, say $\mathrm{d}^2 u$, hence integrating over time still results in an infinitesimal change $t\,\mathrm{d}^2 u$ from which we would conclude that it is impossible for bodies to change temperature.

Laplace proposes that 'the preceding difficulty relating to heat will be eliminated by extending its action beyond contact'. Drawing on the known phenomenon of heat radiation in air, he suggests that

> it seems natural to admit this radiation of heat within dense bodies; however, the internal radiant heat is completely intercepted by molecules very close to the one that heats them, and whose heating action then extends only to a very small distance.

To this end, he introduces a function $\phi(s)$ which expresses how much heat is communicated over a distance s. Its exact form is unknown, but we know $\phi(s)$ becomes zero when s becomes a sensible distance. He then states the integral that expresses the heat gained by the molecule at $x = 0$ during the instant dt:

$$K \int \Big(u(x-s) - 2u(x) + u(x+s) \Big) \phi(s)\, \mathrm{d}s. \tag{30}$$

Unlike Laplace, I made the dependence of $u_{,}, u$ and u' on the spatial coordinate x and relative distance s explicit, using $u(x-s)$ and $u(x+s)$.

adding that we should integrate from $s=0$ to $s=\infty$.

Subsequently, we expand $u(x-s)$ and $u(x+s)$ into a Taylor series around x with respect to the distance s, leading to

$$u(x-s)-2u(x)+u(x+s)=s^2\frac{\mathrm{d}^2u}{\mathrm{d}x^2}+\mathcal{O}(s^4)$$

Laplace simply writes '...' where I put $\mathcal{O}(s^4)$.

Suppressing the higher order terms, Eq. (30) becomes

$$K\frac{\mathrm{d}^2u}{\mathrm{d}x^2}\int_0^\infty s^2\phi(s)\,\mathrm{d}s.$$

The integral in this expression is, like K, a quantity solely dependent on the nature of the body. Designating $a=K\int_0^\infty s^2\phi(s)\,\mathrm{d}s$, one obtains the heat equation in one dimension

$$\frac{\mathrm{d}u}{\mathrm{d}t}=a\frac{\mathrm{d}^2u}{\mathrm{d}x^2}.$$

Laplace does not solve the equation in this note. Its primary purpose is not to develop a mathematical theory of heat but to indicate what the foundations of such a theory are. He concludes the note with the words:

> The rest is a matter of analysis, and becomes foreign to the object of this Note, in which I have sought to establish that the phenomena of nature are in the last analysis reduced to actions *ad distans* from molecule to molecule, and that the consideration of these actions must serve as the basis of the mathematical theory of these phenomena.

Laplace had a broader purpose with this note beyond the particular mathematics of heat: the promotion of what Robert Fox has called 'the Laplacian program'. This program characterized the later part of his career, the first part being defined by his work in celestial mechanics. Laplace sought to extend, by analogy, the essential perspective in astronomy—the planets as point-masses governed by central-force interactions—and the successes of their mathematical theory, to the microscopic realm: forces of attraction and repulsion between particles of matter (molecules) amenable to mathematical analysis. In other words, he posited bodies to consist of molecules attracting or repelling each other with laws analogous to those of gravity. The purpose of physics is then to figure out these laws and develop their mathematical theory. This perspective has its origins in remarks made by Newton in the *Principia* and *Opticks*:

See Fox 1974.

> For many things lead me to have a suspicion that all phenomena may depend on certain forces by which the particles of bodies, by causes not yet known, either are impelled toward one another and cohere in regular figures, or are repelled from one another and recede. Since these forces are unknown, philosophers have hitherto made trial of nature in vain. But I hope that the principles set down here will shed some light on either this mode of philosophizing or some truer one.

Newton 1999, pp.28–9; cf. Newton 1952, Query 31.

Laplace took Newton's more speculative remark, which still leaves room for 'some truer mode of philosophizing', and turned the idea into a research program. The application of this program to the phenomenon of heat includes the *caloric hypothesis*: the idea that heat is a weightless material fluid. Its particles repel one another (leading to the elastic effects of heat) but are attracted to ordinary matter through short-range central forces. This reduces thermal phenomena to molecular interactions governed by the action *ad distans* Laplace describes in the passage quoted above. Poisson, likewise, wrote in his 1835 book:

> The hypothesis that makes the phenomena of heat depend on the undulations of a stagnant fluid has led, until now, to no precise result conforming to experience; this is why I will adopt in this work the much more fruitful theory, which attributes these phenomena to an imponderable matter, contained in the parts of all bodies as small as one may wish, and being able to detach itself from them and pass from one part to another, which causes the quantity of this substance enclosed in each part to vary with time. This matter is called caloric; it is also named matter of heat, or simply heat.

Poisson 1835, §1

However, Laplace's usage of 'molecule' in his note deserves more scrutiny. Presumably, 'molecules' refer to discrete microscopic particles; yet in Eq. (30), we do not sum over all molecules but integrate over infinitesimal slices ds. Laplace seems to equivocate between actual physical particles separated by spaces of air and a more generic usage of molecule as 'a tiny part of a continuous body'. Given that Laplace did hold the former physical perspective, there is a gap between the physical hypothesis and his mathematical derivation. Whether Laplace himself was unaware of this or whether the note was only supposed to indicate a method, instead of giving a full account, would require a more systematic study of Laplace's ideas about 'molecule'. In any case, it was left for Poisson to elaborate on Laplace's note. To fully appreciate what distinguishes Fourier's method from the Laplacian one, we will consider aspects of Poisson's derivation in his 1835 book.

We will see Fourier use 'molecule' in this more generic way, see Chapter 2 section I.

Laplace lead us to consider two scales in his derivation: the macroscopic scale of bodies and the microscopic scale of molecules. Poisson adds a third *mesoscopic* scale to this by defining a material part m 'whose dimensions are insensible, and which contains nevertheless an immense number of molecules'. He elaborates on why we need this scale:

> The exchange of heat between m and another similar part m', will result from the emission and absorption by all their molecules. But if one considered in isolation a molecule of m and a molecule of m', this exchange would present nothing regular that one could submit to calculation: at certain intervals of time, the molecule of m might not emit heat towards the molecule of m', or the latter might send nothing to the first; the opposite would take place at other epochs. At a same instant, the exchanges of heat could be very different between the molecule which corresponds to point M and the molecules situated, in various directions, at equal distances from point M; and finally the exchanges would also vary without any regularity, in passing from this point M to another point situated at an insensible distance from M. The extremely large numbers of molecules of which the material parts, such as m and m', are composed, produce, in all these respects, the regularity indispensable in the exchanges of heat, so that one can calculate the variations of temperature which result from them, and express the temperature corresponding to any point M, as a function of its three coordinates and of the time t; which is the general object of the problem we will have to solve.

Cf. the modern distinction between a Representative Volume Element and a Statistical Volume Element.

Poisson 1835, §7

In other words, recognizing the possible random variability of the microscopic scale, he emphasises the need to consider a scale at which these effects homogenize. Using the microscopic relations to define relations between mesoscopic 'parts of matter', Poisson deduces that the heat exchanged by the parts m and m' during $\mathrm{d}t$ is

$$\frac{pmm'}{4\pi r^2}(q'\Pi - q\Pi').$$

Instead of following his deduction, we simply state what the various terms mean, which gives an indication as to what kind of reasoning went into the deduction.

Symbol	Meaning
m, m'	Masses of two material parts of insensible dimension
r	Distance between them
p	Absorption factor: fraction of heat surviving passage through the intervening matter
Π, Π'	Emitting powers of m, m' per unit mass per unit time
q, q'	Absorbing powers of m, m'

Only now does Poisson invoke Newton's law as a linearisation, replacing $q'\Pi - q\Pi'$ with $u' - u$. Absorbing the other factors into a function ϕ which depends on the nature of the material and summing over all m' in a radius of activity around m, we get the discrete sum

$$\sum \frac{\phi}{r'^2}(u' - u)v'$$

with v' and r' being respectively the volume of the material part m' and the distance between m and m'. Poisson is working in three dimensions so that this discrete sum is subsequently approximated by an integral

$$\iiint \frac{\phi(r')}{r'^2}\Big(u'(x',y',z') - u\Big)\,\mathrm{d}v' \tag{31}$$

$\mathrm{d}v' = \mathrm{d}x' \times \mathrm{d}y' \times \mathrm{d}z'$ and $r'^2 = x'^2 + y'^2 + z'^2$.

by invoking the Euler-Maclaurin formula. A power series expansion of u' with a suppression of its higher order terms, plus a few pages of multivariable integral calculus (with further simplifications), gives the heat equation in three dimensions:

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \kappa\left(\frac{\mathrm{d}^2u}{\mathrm{d}x^2} + \frac{\mathrm{d}^2u}{\mathrm{d}y^2} + \frac{\mathrm{d}^2u}{\mathrm{d}z^2}\right),$$

where I collected the constants into a single κ. The heat equation for a spherical body (where x denotes the distance to the center and without angular dependence) is also derived from Eq. (31).

See Poisson 1835, §137 for the spherical case.

★ III.2 INTEGRATING THE HEAT EQUATION: POISSON'S METHOD

We now turn to Poisson's method for solving the heat equation. Recall that our primary purpose in the first two chapters is to give an account of Fourier's ideas in the *Théorie* and that we are describing Poisson's work because Fourier is contrasting himself with it. However, there are various elements in Poisson's mature 1835 book—published after Fourier's death—that make his method closer to that of Fourier; in particular, his expansion of the sought temperature function $u(x,t)$ in powers of e^{-t}. I will therefore set these aside and present a synthesis of his 1835 book and a 1815 mémoir, since Fourier is addressing this mémoir directly. The deeper justification for this is that there is a genuinely different approach to solving the problem that Fourier is contrasting himself with, the method Laplace used in the problems of sections I and II: first derive, for the sought function, an integral representation in finite terms and involving unknown arbitrary functions, and then secondarily derive a series expansion from the integral. This method can be used for the heat equation too. Consequently, it is most instructive to consider the two archetypes pure, while keeping in mind that there are elements in Poisson's work that make him more similar to Fourier.

Poisson presents his solution in the 1835 book as part of the chapter '*Digression sur les intégrales des équations aux différences partielles*'. The chapter presents a general method for solving (linear) PDEs and we quote a significant part of it:

> Let u be a function of any number of independent variables $t, x, y, z, \ldots$ which must satisfy a partial differential equation of order n, represented by L. Whatever the unknown value of u may be, we can conceive it expanded according to the ascending powers of one of these variables, diminished by a particular value of this same variable; for if we call h a particular value of t, for example, and put $h+t-h$ instead of t in the function u, we can always expand it by Taylor's theorem, or otherwise, in an ordered series according to the ascending powers of $t-h$.
>
> More generally, if we take for θ a quantity dependent on one or more of the variables t, x, y, z, etc., we can suppose the value of u expanded according to the ascending powers of θ. Let therefore
>
> $$u = P\theta^{\alpha} + Q\theta^{\beta} + R\theta^{\gamma} + \ldots; \tag{32}$$
>
> the coefficients P, Q, R, etc., being unknown functions each of which depends on one less independent variable than u, and the exponents α, β, γ, etc., forming an increasing sequence of constant quantities. This being posited, if we substitute this value of u in the equation $L = 0$, order its first member according to the powers of θ, and then equate separately to zero the coefficient of each of the terms of the series which will result, we will thus have an infinite sequence of differential or partial differential equations, each of which will contain one independent variable less than $L = 0$; and if we manage to obtain the most general values of $\alpha, \beta, \gamma \ldots$ and $P, Q, R \ldots$, which satisfy this sequence of equations, the series (32) will also be the most general value of u that will satisfy the given equation $L = 0$. Depending on the quantity θ that we choose, we will thus have different expressions of u in series which will be, under equivalent forms, the complete integral of $L = 0$, so that if this integral were known in finite form, each of these series would be an expansion of it that could replace it. Poisson 1835, §71

Poisson subsequently applies this general procedure to the one-dimensional heat equation Poisson 1835, §§72–74

$$\frac{\mathrm{d}u}{\mathrm{d}t} = a\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}. \tag{33}$$

Taking 'h une constante à laquelle on n'attribue aucune valeur déterminée', he considers $\theta = t - h$ so that the coefficients Eq. (32) are functions of x, since u is a function of t and x. Filling the series into Eq. (33) he first deduces

$$\alpha = 0, \quad \beta = 1, \quad \gamma = 2, \quad \ldots$$

and subsequently derives

$$u(x,t) = \varphi(x) + a(t-h)\frac{\mathrm{d}^2\varphi(x)}{\mathrm{d}x^2} + \frac{a^2(t-h)^2}{2!}\frac{\mathrm{d}^4\varphi(x)}{\mathrm{d}x^4} + \ldots \tag{34}$$

for an arbitrary function $\varphi(x)$. 'l'on peut regarder comme la valeur la plus générale de u tant que l'on n'y détermine pas la constante h'. The reason for the indeterminate h is 'functions that can only be expanded according to the descending powers of the variable; and the function u being unknown, it can be of this kind'.

$\frac{1}{t} = \frac{1}{h+(t-h)} = \sum_{n=0}^{\infty}(-1)^n\frac{(t-h)^n}{h^{n+1}}$, but expanding $\frac{1}{t}$ in pure powers of t is impossible.

To obtain a finite equation for u, Poisson reduces Eq. (34) to a definite integral. He recalls (and re-proves)

$$\int_{-\infty}^{+\infty} e^{-g\omega^2}\,\mathrm{d}\omega = \sqrt{\frac{\pi}{g}}, \tag{35}$$

and 'differentiating n times in succession with respect to g, and then making $g = 1$,' he deduces

$$\int_{-\infty}^{+\infty} \omega^{2n}e^{-g\omega^2}\,\mathrm{d}\omega = \frac{(2n)!}{2^n n!} \quad \text{and} \quad \int_{-\infty}^{+\infty} \omega^{2n+1}e^{-g\omega^2}\,\mathrm{d}\omega = 0.$$

He uses these to rewrite Eq. (34) as

$$u(x,t) = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty}\Big\{\varphi(x) + 2\omega\sqrt{a(t-h)}\frac{\mathrm{d}\varphi(x)}{\mathrm{d}x} + \frac{\left(2\omega\sqrt{a(t-h)}\right)^2}{2!}\frac{\mathrm{d}^2\varphi(x)}{\mathrm{d}x^2} + \frac{\left(2\omega\sqrt{a(t-h)}\right)^3}{3!}\frac{\mathrm{d}^3\varphi(x)}{\mathrm{d}x^3} + \ldots\Big\}e^{-\omega^2}\,\mathrm{d}\omega, \tag{36}$$

Note that the odd terms of Eq. (36) all equal zero, they are added purely to deduce Eq. (37).

which by Taylor's theorem reduces to

$$u(x,t) = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty}\varphi(x + 2\omega\sqrt{a(t-h)})e^{-\omega^2}\,\mathrm{d}\omega. \tag{37}$$

For $h = 0$, Poisson argues that

> this equation is, in finite and simplest form, the complete integral of equation (33). It contains, as we see, only a single arbitrary function, which will be determined immediately from the value of u relative to $t = 0$.

He immediately qualifies this statement: 'this form supposes that this value of u, which will be that of $\varphi(x)$, grows with the variable x in a lesser ratio than e^{x^2}'.

Poisson spends some time studying Eq. (37). Through a change of variables he deduces a different form:

Poisson 1835, §§74–76

$$u(x,t) = \int_{-\infty}^{+\infty}\frac{1}{2\sqrt{\pi a(t-h)}}e^{-\frac{(\omega-x)^2}{4a(t-h)}}\varphi(\omega)\,\mathrm{d}\omega. \tag{38}$$

He also repeats similar procedures for power expansions of u in $x - h$ and e^x, showing that they lead to the same Eq. (37).

Subsequently, he extends 'without difficulty' the argument to a function u of an arbitrary number of variables $t, x, y, z\dots$ governed by

$$\frac{\mathrm{d}u}{\mathrm{d}t} = a\frac{\mathrm{d}^2u}{\mathrm{d}x^2} + b\frac{\mathrm{d}^2u}{\mathrm{d}y^2} + c\frac{\mathrm{d}^2u}{\mathrm{d}z^2} + \dots,$$

obtaining

$$u = \int\cdots\int \varphi\left(x + 2\omega\sqrt{a(t-h)}, y + 2\omega'\sqrt{b(t-h')}, \dots\right) e^{-\omega^2-\omega'^2-\cdots}\,\mathrm{d}\omega\,\mathrm{d}\omega'\dots.$$

Afterwards, he turns to different PDEs to show that the same procedure can be used to obtain their integrals 'sous forme finie'. This deserves mention since it emphasises how obtaining such an integral is the goal according to Poisson, it is what it means to 'integrate' the PDE.

Poisson 1835, §§76–82

We will elaborate on this point after presenting the second part of Poisson's solution.

For the second part of the solution—imposing the special boundary conditions—we will follow Poisson's 1815 mémoir and consider the one-dimensional case. Suppose we have an IC and we set the temperature to zero at the two boundaries. Imposing an IC is trivial, since by setting $t=0$ in Eq. (37) we see that φ simply is the IC. However, it is not obvious how to apply BCs. To this end, Poisson uses what today is known as 'the method of images'. For the sake of clarity, I am presenting my condensed summary of Poisson's essential steps, using the notation of Poisson's 1835 book; a more detailed description of Poisson's arugment is found in Arnold 1983.

Poisson 1815, §29

The IC only specifies the temperature within the boundaries, say, from $-l$ to l. Hence, we still have additional freedom to define φ outside this interval. To this end, let us consider $\varphi(x)$ as a sum of functions

$$\begin{aligned}\varphi(x) = \phi_0(x) + \phi_1(2l-x) + \phi_2(4l-x) + \phi_3(6l-x) + \dots\\ + \phi_1(-2l-x) + \phi_2(-4l-x) + \phi_{-3}(-6l-x)\end{aligned}$$

where each ϕ_n is nonzero only on $2(n-\frac{1}{2})l < x < 2(n+\frac{1}{2})l$. Hence, the IC only applies to $\phi := \phi_0$ so that we are free to choose the other ϕ_i arbitrarily. We will use this freedom to satisfy the BC for all values of t. In particular, it turns out that we can simply put

$$\begin{aligned}\varphi(x) &= \phi(x) - \phi(+2l-x) + \phi(+4l-x) - \phi(+6l-x) + \phi(+8l-x)\dots\\ &\quad - \phi(-2l-x) + \phi(-4l-x) - \phi(-6l-x) + \phi(-8l-x)\dots\\ &= \sum_{n=-\infty}^{\infty} \phi(4nl+x) - \phi(4nl+2l-x)\end{aligned}$$

Substituting this into the general equation gives

$$\begin{aligned}u(x,t) = \frac{1}{\sqrt{\pi}}\sum_{n=0}^{\infty}\int_{-\infty}^{+\infty}\Big[\varphi(x+4nl+2a\omega\sqrt{t}) - \varphi(2l-x+4nl+2a\omega\sqrt{t})\\ + \varphi(x-4l-4nl+2a\omega\sqrt{t}) - \varphi(-x-2l-4nl+2a\omega\sqrt{t})\Big]e^{-\omega^2}\,\mathrm{d}\omega.\end{aligned}$$

Poisson then makes various substitutions to derive

$$u(x,t) = \sum_{n=0}^{\infty} a_n e^{-\frac{a^2t\pi^2n^2}{l^2}}\sin\left(\frac{n\pi x}{l}\right) + b_n e^{-\frac{a^2t\pi^2(2n+1)^2}{4l^2}}\cos\left(\frac{(2n+1)\pi x}{2l}\right)$$

with

$$a_n = \int_{-l}^{+l} \phi(y) \sin\left(\frac{n\pi y}{l}\right) \mathrm{d}y \quad \text{and} \quad b_n = \int_{-l}^{+l} \phi(y) \cos\left(\frac{(2n+1)\pi y}{l}\right) \mathrm{d}y.$$

This is the solution to the one-dimensional bar problem in the form of a Fourier series. The form of the solution as a series is thus derived from the integral, which is considered the general solution.

Poisson's method for solving the heat equation is essentially similar to Laplace's method for both the gravitational potential of section I as well as the urn problem of section II. It involves two steps:

(1) The general solution: integrate the PDE to derive an integral expression for u involving one or more arbitrary functions.

(2) The particular solution: derive what the arbitrary function(s) in the general integral solution should be to satisfy particular boundary and initial conditions, possibly by expanding functions into series.

Cf. the integral equation for the gravitational potential and Eq. (21).

This method will be contrasted with Fourier's approach—solving problems with particular BC by looking for particular 'simple modes' first and subsequently arguing that a sum of 'simple modes' can describe an arbitrary function—in Chapter 2 section II.

The integral form obtained in step (1) can be studied on its own, for example under changes of variables. Poisson seems to think that the PDE without other conditions is the most general problem and the integral its most general solution. The special conditions impose particular restrictions on the general solution and should be considered secondarily. Poisson obtains a series solution in step (2), as a *consequence* from the general solution to the PDE. His criticism of Fourier's approach is that he did not derive his series solution from the general integral solution:

> the only way to remove all doubt and to conserve the mathematical certainty of the solution is not to suppose in advance such a form for the unknown, but rather to deduce it from the general integral by a series of direct and rigorous transformations.

Cited and translated in Arnold 1983, p.314. This perspective is absent in the 1835 book.

In 1835, he sidesteps the transformation of the integral into the series—a move Fourier criticized in the *Théorie*—and instead directly expands u into powers of e^{-t} (which, by assumption, is always possible). The coefficients of this series are consequently trigonometric functions and the decomposition of an arbitrary function into a trigonometric series follows by letting $t = 0$. The view present in both works is that the decomposition of arbitrary functions into a trigonometric series can and should be considered a special case of the solution to the heat equation—the case that $t = 0$—rather than a purely mathematical fact of 'general analysis' that is proven independent of the physical phenomenon, and which is *applied* to solve the heat equation. Poisson acknowledges that the decomposition then depends on the fact that 'the problem is susceptible of a solution'. That premise either requires us to postulate directly the existence of a solution to the mathematical problem, or to derive it from the existence of the physical phenomenon by arguing that the mathematical equations describe it accurately. If the latter justification

See subsection II.4.

See Arnold 1983, p.313 for why the proof involving setting $t = 0$ is fallacious.

is chosen, then the mathematical argument hinges on how accurately the mathematical equations describe the physical phenomenon. In section I we saw Laplace similarly derive properties of spherical harmonics from the physical equation of gravity at the boundary of the spheroid.

All in all, Poisson's criticism does not center on *whether or not* we can decompose functions into the series which show up in solving the heat equation—'although we cannot always directly demonstrate these various formulas, there can nevertheless remain no doubt as to their exactitude'—it centers on the fact that he does not believe that these formuae have been demonstrated directly in all cases and he suggests an alternative way to prove them generally. He acknowledges that Fourier's approach 'can be regarded as complete in the cases where it has been demonstrated that the value of the unknown which corresponds to $t=0$ can represent any function'. He seems to think Fourier has done so for the case of developing arbitrary functions as a trigonometric series with integer frequencies, but not for the case of transcendental frequencies (corresponding to the case of convective boundary conditions).

Poisson 1835, §86

Poisson 1835, §88

Chapter 2

Jean-Baptiste Joseph Fourier

Fourier! with solemn and profound delight,
Joy born of awe, but kindling momently
To an intense and thrilling ecstasy,
I gaze upon thy glory and grow bright...

— Sir William Rowan Hamilton

Graves 1882, p. 596

Fourier's career is a heroic pursuit of mathematics in the face of political interruptions that reality threw his way. Born on March 21, 1768 in Auxerre, he fell in love with mathematics at the age of thirteen, collecting candle-ends by day so that he could study by night in a storeroom of the *École Royale Militaire* at Auxerre. A life of teaching and research seems to have been what he sought, but when the French rose against the monarchy, it drew him in. 'During the early years of the Revolution he seems to have behaved honourably and therefore was imprisoned by both sides' according to Ivor Grattan-Guinness. A few years of mathematics followed, as a student at the newly founded *École Normale* and then as a teacher at the *École Polytechnique*, until Napoleon pressed him into his Egyptian expedition as a scientific advisor. Back from Egypt, he was made prefect of Isère, cut off from the capital of science that was Paris. Yet he dedicated his late hours to mathematics, first repeating experiments of his predecessors and subsequently developing his own mathematical theory of heat grounded in the results of those experiments. After Napoleon's first fall he returned to Paris, but Napoleon came back and Fourier was sent away to govern the Rhône. With the final defeat of Napoleon in 1815, Fourier's name was tainted by association with the fallen regime. Stripped of his pension, he climbed his way back through the Parisian scientific ranks, until the year 1822, when he published his *Théorie analytique de la chaleur* and was elected permanent secretary of the *Académie des sciences*. The last eight years of his life were fully devoted to science, at last achieving the career he sought.

Grattan-Guinness 1990, p.584

For most of his life Fourier was thus isolated from the research community. This isolation was as much physical as it was intellectual. His 1807 memoir on heat, already containing most of the ideas found in the *Théorie*, was met with serious objections, chiefly from Lagrange and

This biographic sketch as well as the following paragraph is primarily based on Herivel 1975, supplemented by Grattan-Guinness and Ravetz 1972.

Laplace over his use of trigonometric series. The commission charged with reporting on it never reported, and the memoir went unpublished. In 1811 the *Institut* set the propagation of heat as its grand prize. Fourier submitted an enlarged but essentially unrevised version of the 1807 memoir and won. However, the commission—Lagrange, Laplace, Malus, Haüy, Legendre—was largely the same set of critics as before, and they described their reservations in the award itself. Fourier apparently protested, but without success. Nor did the prize lead to the publication of his work: the essay stayed in manuscript until 1822, the year of the *Théorie*. The reservations are worth quoting:

Herivel 1975, p. 100

Herivel 1975, p. 103

> the manner in which the author arrives at these equations is not exempt of difficulties and that his analysis to integrate them still leaves something to be desired on the score of generality and even rigor.

Herivel 1975, p. 103

We seek to give an account of Fourier's mature understanding of the two things about which this quote expresses reservations: the manner in which the heat equation is derived, motivating a discussion of the broader method of the *Théorie* (section I); and the analysis which solves the equations thus derived (section II). The latter includes the issue of the generality of Fourier series. Given the scope of this topic—including the definition of 'function'—it deserves its own section (section III).

I Foundations of Fourier's *Théorie*

I will discuss two fundamental parts of Fourier's derivation: the basic physical principle that forms the foundation for his theory (subsection I.1) and the argument by which he derives what is today called 'Fourier's law' from this basic principle (subsection I.2). Based on these, I will then give a philosophical account of Fourier's method (subsection I.3).

I.1 The principle of the communication of heat versus the mode of action of heat

The principle which serves as the physical foundation for Fourier's mathematical theory of heat is what he calls *the principle of the communication of heat* (PCH). In the summary at the end of the *Théorie*, he formulates it thus:

Théorie Ch. 1 Sec. III.

> If two molecules of the same body are extremely near, and are at unequal temperatures, that which is the most heated communicates directly to the other during one instant a certain quantity of heat; which quantity is proportional to the extremely small difference of the temperatures.

Théorie §429. All article numbers subsequently listed refer to the *Théorie*, unless indicated otherwise.

In the text he also translates it into analytical symbols, writing:

> Denoting by v and v' the temperatures of two equal molecules m and n, by p, their extremely small distance, and by $\mathrm{d}t$, the infinitely small duration of the instant, the quantity of heat which m receives from m' during this instant will be expressed by $(v'-v)\phi(p)\,\mathrm{d}t$. We denote by $\phi(p)$ a certain function of the distance p which, in solid bodies and in liquids, becomes null when p has a sensible magnitude. The function is the same for every point of the same given substance; it varies with the nature of the substance.

§52

This principle is deeply related to Newton's law of cooling and Fourier at times refers to it as Newton's principle. However this principle is, strictly speaking, more general. Newton's principle concerns the behavior of discrete (or contiguous) bodies, whereas Fourier's statement pertains to *molecules*. 'Molecule', in this context, is simply a generic name for a tiny part of a continuous body considered in isolation from the rest; Fourier uses it interchangeably with 'material point'. Most importantly, it is a *mathematical* rather than a *physical* concept and entails no view about the microscopic nature of matter. Using 'molecule' this way does not imply that matter consists of discrete particles, it is merely a perspective on the continuous: bodies can be divided into (arbitrarily small) parts and such parts can be considered in isolation. In this respect, it is closest to Poisson's 'part of matter'. According to Fourier, the PCH 'is confirmed by multiple observations and [is] accepted by all physicists', which presumably explains why he doesn't spell out the experiments and argument that establish the principle. What was controversial is how it can be used to derive the heat equation and whether or not that requires Laplace's 'molecular hypothesis'—a statement about *physical* molecules.

See Fourier n.d.(b), fol.152–58. See also Fourier 1810a in Fourier 1980 p. 33.

'Material point' is used in §57 and §429 n.9.

See p. 25 for Poisson on 'parts of matter'.

Fourier 1810a in Fourier 1980 pp.27–35. See also §15.

Let us interject that there are two essentially independent physical ideas which the PCH rests on. Having divided a continuous body into molecules $M_1, M_2, \dots, M_n$, we can consider the net amount of heat communicated to a particular molecule M_k during an interval T to be a quantity H_k which depends on various characteristics of the molecules $M_1, M_2, M_3, \dots, M_n$. The first idea is that, if T is some infinitesimal time interval $\mathrm{d}t$, then $H_k = \sum_i h(M_k, M_i)\,\mathrm{d}t$. In other words, the total effect of the molecules is the sum of their independent effects $h(M_k, M_i)\,\mathrm{d}t$. This is not self-evident, it is clearly false for finite time intervals: given three nearby molecules a, b and c, how much heat molecule a communicates to b is affected by how much c communicates to either. Secondly, there is the additional idea that h has a specific form: $h(M_k, M_i) = (v_i - v_k)\phi(p)$, where $v_i - v_k$ is the temperature difference of the molecules and $\phi(p)$ is some unknown function of the distance p between the molecules. In other words, the net amount of heat M_i communicates to M_k depends in a specific way on: (a) the temperature difference between them (and not their absolute temperature); (b) the material, which determines ϕ; and (c) the distance between the molecules.

From the PCH, Fourier will deduce what is today known as Fourier's law and from Fourier's law he will derive the heat equation. Before turning to those derivations, we will elaborate on the physical foundation of Fourier's mathematical theory. Though the most important, the PCH is not the only thing that Fourier discusses before turning to pure mathematics. In Section II of Chapter 1, preceding the section on the PCH, he considers various basic characteristics of heat: he introduces the properties of specific heat, interior and exterior conductivity, all three of which define constants entering into the equations; he discusses three different mechanisms of heat transfer: conduction, convection and radiation, and names the tendency of heat towards equilibrium distribution. The latter is a universal effect which is independent of contact, it also

Fourier does not use the terms conduction and convection.

applies to bodies radiating in a vacuum. In the final part of Section II he turns to some basic aspects of radiant heat. He also, famously, does not take a position on the nature of heat: 'of the nature of heat only uncertain hypotheses could be formed'. However, Section II *does* include a discussion of what he calls the *mode of action* of heat. The distinction between the nature and mode of action of heat deserves more of our attention.

§§40–56

§21

I am not aware of any discussion of this distinction in the literature on Fourier.

Fourier initially distinguishes heat radiation from the other two mechanisms of heat transfer, but in the final part of Section II he proposes a more sophisticated perspective: *heat transfer by contact is simply radiation on a microscopic scale.* In other words, the universal *mode of action* of heat is radiation: 'heat acts in the same manner in a vacuum, in elastic fluids, and in liquid or solid masses, it is propagated only by way of radiation, but its sensible effects differ according to the nature of bodies'. He considers the mode of action of heat to be like that of light: 'molecules separated from one another reciprocally communicate, across empty space, their rays of heat, just as shining bodies transmit their light'. The view of heat as universally acting by radiation is not unique to Fourier, the Laplacians held the same view. What is distinctive to Fourier is (a) that he takes a position on the *mode of action* of heat without taking a position on the *nature* of heat, unlike the Laplacians who held that caloric matter was radiating and (b) his insistence that the subsequent mathematical theory—in particular the PCH—does *not* depend on this physical perspective: 'it is not necessary to consider the phenomena under this aspect in order to establish the theory of heat'.

§52. See also Fourier 1810a in Fourier 1980 p. 33.

§40

§56. See also Fourier 1810b in Fourier 1980 p. 38.

The distinction between the nature and mode of action of heat seems to be a mature perspective absent from Fourier's earlier drafts. Consider one such draft, probably from 1817:

> For us it seems very important to us not to give the principle of the communication of heat any hypothetical extension, and we think that this principle is sufficient to establish the mathematical theory because the equations are proved in the clearest and most rigorous way without having to examine whether propagation is exerted by radiation in the interior of solids, if it consists or not in the emission of a proper matter which the molecules send to each other reciprocally, or whether it results, as sound, from the vibrations of an elastic medium. It is always preferable to limit oneself to enunciate the general fact that the observations indicate, which is none other than the preceding principle [the PCH]. It is thus shown that the mathematical theory of heat is independent of any physical hypothesis; in fact, the laws to which propagation is subject are admitted by all physicists, in spite of the extreme diversity of their sentiments on the nature and mode of its action.

Fourier n.d.(b), fol.158 [1]; the dating is taken from Charbonneau 1994 (henceforth cited as Charbonneau, Catalogue). This particular note is written up neatly by a copyist, unlike the one below.

Regarding the above two points, observe on the one hand, his insistence on separating the (universally accepted) *mathematical* PCH from the (more controversial) *physical* aspects of heat; and on the other hand, the three positions he names on 'the nature and mode of action' of heat: radiation, emission of a proper matter, and vibration. In the *Théorie*, he *separates* the nature and the mode of action of heat; he suggests a view on the mode of action of heat, but he judges the evidence about heat's nature to be inconclusive. At the end of the *Théorie* he mentions the *two* classical positions on the nature of heat: 'in whatever manner we please to imagine the nature of this element, whether we regard it as a distinct material thing which passes from one part of space to another, or whether we make heat consist simply in the transfer of motion …', §432 i.e. he considers the caloric or kinetic theory. But this distinction now pertains purely to the nature of heat, heat being subject to a *single* mode of action in both cases: radiation, either of 'a distinct material thing' or of 'motion'.

We might have evidence of the actual note where Fourier first introduced this distinction. We can see him write the following in a note, probably from 1818:

> We have not spoken hitherto of the physical nature of heat, for that knowledge is in no way necessary in order to found the mathematical theory. But one cannot connect all the phenomena with one another without representing to oneself under a distinct form ~~the nature and~~ the action of the principle. It would be a great error to compare heat to an elastic fluid such as air or the gases. Matter exists in the world under four different forms: solid substances, liquid masses, aeriform fluids or gases, and radiant substances such as light, which traverse all the points of space with an immense velocity, so that there is no point which is not the centre of a sphere of light. This mode of universal being is the one proper to the principle of heat.

Fourier n.d.(b), fol.212–214 [2]; strikethrough in the original.

He originally wrote 'the nature and the action of the principle' and then *struck out* 'the nature', retaining only the need to say something about its mode of action. We also see him reject the view of heat as 'an elastic fluid' (i.e. the caloric theory of heat), a view which is softened in the final *Théorie*: 'the effects of heat can by no means be compared with those of an elastic fluid whose molecules are at rest.' He rejects *static* §52 caloric theories in which the absence of temperature changes implies the absence of heat transfer, but leaves room for dynamic caloric theories where heat is continuously emitted and absorbed by all molecules and the lack of temperature changes merely implies an equilibrium (like the Laplacians held.) Lastly, Fourier also floats the curious idea that heat and light form a *fourth* of form of matter, distinct from solids, gasses and liquids. This perspective is absent in the *Théorie* where the caloric and kinetic theory are suggested as possible views on the nature of heat.

The composition of the above notes coincides chronologically with Fresnel's publications on light. From about 1815 onwards, Fresnel started publishing his work on light, launching a serious attack on the Laplacian corposcular theory of light and putting forward an alternative wave theory of light. A very favourable report—co-written by Fourier, Ampère and Arago—on Fresnel's 1821 definitive paper about double refraction includes the following section:

> Mr. Fresnel presents his theoretical ideas on the particular type of undulations which, according to him, constitute light; time has not yet allowed us to examine them with all the necessary attention. It would be impossible for us to express a definitive opinion on this subject today: the Commission may return to it at another time.

Fourier et al. 1822

The report itself was published in 1822, the same year Fourier's *Théorie* was published. Given Fourier's view that heat and light are similar phenomena, the work of Fresnel—and the expressed agnosticism in this note—should be considered highly significant in relation to Fourier's thinking about the nature of heat. It might explain why the perspective of four forms of matter is absent in the *Théorie* and it is not obvious who influenced whom at what time and about what idea. A better understanding of Fourier's views would require a precise textual analysis of Fourier's drafts and his interaction with Fresnels work. They were both active researchers in Paris at this time and their interactions were certainly not limited to reading each other's formal publications. Such an analysis would be a whole project on its own, impossible in the context of this thesis. Moreover, the necessary evidence to create a precise timeline of Fourier's views might not be extant.

Fourier's discussion of the physical aspects of heat in the *Théorie*, even apart from the manuscripts, call into question the conventional view that Fourier was a positivist who thought that the nature of heat was beyond the realm of science. We will consider Fourier's alleged positivism in subsection I.3.

I.2 The heat flux and Fourier's law

From the PCH, Fourier deduces what is today known as Fourier's law by a clever and creative argument. Unlike Poisson, who calculated the effects that all surrounding molecules have on one particular molecule by brute force, Fourier first turns to a linear temperature distribution between two infinite planes $x = A$ and $x = B$ separated by a distance e and held at temperatures a and b respectively. Hence $u(x) = a + \frac{b-a}{e}x$. His subsequent argument is as follows. Consider two planes $x = A'$, $x = B'$ in between and parallel to A and B. Take any two molecules m, m' on different sides of A'. Any such pair stands in one-to-one relationship with two molecules n, n' on different sides of B' (through the translation $x \mapsto x + B' - A'$). The amount of heat communicated by n to n' equals that communicated by m to m' because of (a) the homogeneity of the body and (b) the linearity of the temperature distribution. Since this holds for *any* pair of molecules, the amount of heat flowing through (any equal portion of) A' and B' is equal. Since the flow in and out of the slice contained by A' and B' is equal, the amount of heat it contains is constant and temperatures will not change, the situation is proven to be a steady-state.

See Fig. 1.

(a) $\phi(p)$ is translation-invariant; *(b)* by definition $x_n - x_{n'} = x_m - x_{m'}$, so that temperature differences between the two pairs of molecules in both cases equals $\frac{b-a}{e}(x_n - x_{n'})$

We can take another perspective on this situation by introducing a new quantity: the heat flux, the amount of heat flowing through a unit of area in a unit of time. We define a unit of this quantity by taking two base planes separated over a unit of distance whose two opposite faces are held at temperatures 0 and 1 respectively, taking a plane parallel to the bases and defining K to be the quantity of heat, measured in some unit of heat, flowing through a unit of this surface in a unit of time; K depends on the nature of the material. Using this quantity, the amount of heat flowing through some area ω of a plane equals $K\omega\frac{b-a}{e} = K\omega\frac{\mathrm{d}u}{\mathrm{d}x}$ and

e.g. 0 and 1 as the freezing and boiling points of water.

e.g. a unit of heat as the heat required to convert 1 kg of melting ice into water.

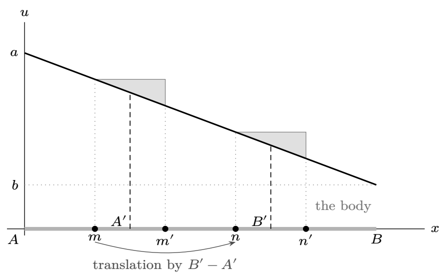


Figure 1. Linear heat flow

we can use this to compare the heat flux in different (one-dimensional) linear temperature distributions for homogenous bodies. In this linear case we have made no approximations at all, but note the suggestive introduction of $\frac{\mathrm{d}u}{\mathrm{d}x}$. Fourier then generalizes the above argument to three dimensions, i.e. a three-dimensional solid with temperature distribution $u = A + ax + by + cz$ and easily derives that the above property of the heat flux holds for the x, y and z directions.

Having spent significant time on the case of linear temperature distributions, we easily obtain the heat flow in a solid with an arbitrary temperature distribution through lineralization. Fourier notes that, by taking a power expansion of the temperature at a point, the solid is locally indistinguishable from one with a linear temperature distribution, so that

'The solid' refers to a body with an arbitrary temperature distribution; 'the prism' refers to one with a linearized temperature distribution.

> the quantity of heat which flows in the solid across the circle ω, during the instant $\mathrm{d}t$, is the same as that which flows in the prism across the same circle; for all the molecules whose actions concur in one effect or the other, have the same temperature in the two solids. Hence, the flow in question, in one solid or the other, is expressed by $-K\frac{\mathrm{d}v}{\mathrm{d}z}\omega\,\mathrm{d}t$. It would be $-K\frac{\mathrm{d}\eta}{\mathrm{d}y}\omega\,\mathrm{d}t$, if the circle ω, whose centre is m, were perpendicular to the axis of y, and $-K\frac{\mathrm{d}v}{\mathrm{d}x}\omega\,\mathrm{d}t$, if this circle were perpendicular to the axis of x. The value of the flow which we have just determined varies in the solid from one point to another, and it varies also with the time. We might imagine it to have, at all the points of a unit of surface, the same value as at the point m, and to preserve this value during unit of time; the flow would then be expressed by $-K\frac{\mathrm{d}v}{\mathrm{d}z}$, it would be $-K\frac{\mathrm{d}v}{\mathrm{d}y}$ in the direction of y, and $-K\frac{\mathrm{d}v}{\mathrm{d}x}$ in that of x. We shall ordinarily employ in calculation this value of the flow thus referred to unit of time and to unit of surface. §98
>
> This theorem serves in general to measure the velocity with which heat tends to traverse a given point of a plane situated in any manner whatever in the interior of a solid whose temperatures vary with the time. §99

Today we condense this result by stating that the heat flux at a point equals $-K\nabla u$; this is the form one would find Fourier's law in a modern textbook.

The heat equation is now obtained easily, for various geometries, by slicing the object into infinitesimal 'molecules' and considering the heat fluxes at their boundaries. For example, for a cube Fourier considers a 'prismatic molecule' $\mathrm{d}x \times \mathrm{d}y \times \mathrm{d}z$ and the fluxes across the square faces. If the amount of heat entering across $\mathrm{d}y \times \mathrm{d}z$ (i.e. perpendicular to x)

equals $K\mathrm{d}y\mathrm{d}z\frac{\mathrm{d}u}{\mathrm{d}x}\,\mathrm{d}t$ then that which escapes across the opposite face equals

$$-K\mathrm{d}y\mathrm{d}z\left(\frac{\mathrm{d}u}{\mathrm{d}x}-\mathrm{d}\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)\right)\mathrm{d}t,$$

their sum is $K\mathrm{d}x\,\mathrm{d}y\mathrm{d}z\frac{\mathrm{d}^2u}{\mathrm{d}x^2}\,\mathrm{d}t$. The same argument can be applied to the y and z direction. If we add these together and

> if the quantity which has just been found be divided by that which is necessary to raise the molecule from the temperature 0 to the temperature 1, the increase of temperature which is effected during the instant $\mathrm{d}t$ will become known. Now, the latter quantity is $CD\mathrm{d}x\,\mathrm{d}y\mathrm{d}z$ for C denotes the capacity of the substance for heat; D its density, and $\mathrm{d}x\,\mathrm{d}y\mathrm{d}z$ the volume of the molecule. The movement of heat in the interior of the solid is therefore expressed by the equation
>
> $$\frac{\mathrm{d}u}{\mathrm{d}t}=\frac{K}{CD}\left(\frac{\mathrm{d}^2u}{\mathrm{d}x^2}+\frac{\mathrm{d}^2u}{\mathrm{d}y^2}+\frac{\mathrm{d}^2u}{\mathrm{d}z^2}\right)$$ §128

A similar argument can be used to derive the heat equation for a sphere (using concentric shells) and an infinitely long cylinder (using concentric rings). Leaving aside angular variations, Fourier obtains:

Sphere with radial coordinate x: $$\frac{\mathrm{d}u}{\mathrm{d}t}=\frac{K}{CD}\left(\frac{\mathrm{d}^2u}{\mathrm{d}x^2}+\frac{2}{x}\frac{\mathrm{d}u}{\mathrm{d}x}\right)$$ §113

Cylinder with radial coordinate x: $$\frac{\mathrm{d}u}{\mathrm{d}t}=\frac{K}{CD}\left(\frac{\mathrm{d}^2u}{\mathrm{d}x^2}+\frac{1}{x}\frac{\mathrm{d}u}{\mathrm{d}x}\right)$$ §119

We conclude this subsection with some observations about the differences in derivation between Fourier and the Laplacians, including their treatment in the secondary literature.

(1) To describe the movement of heat, Fourier turns his eyes to the *simplest* case first: a linear temperature distribution. In this context he forms a new concept—the heat flux—and then reduces the more complex case to this simpler case by the basic principle of analysis: linearization. Such an approach of starting with a simple problem and reducing the more complicated problem to the simple one, contrasts with Poisson's approach. His derivation starts from the most *general* problem: some process described by molecular interactions with all coefficients at first taken to be variable. In this complicated problem, he then introduces (reasonable) simplifications insofar as they are necessary to derive the heat equation.

Cf. Poisson's comments at the end of Poisson 1835, Ch.4.

(2) Core to Fourier's derivation is the formation and use of the concept of a *heat flux*; it is the middle term between the PCH and the heat equation. Fourier is self-conscious about its role:

> This notion of the flow is fundamental; in so far as we have not acquired it, we cannot form an exact idea of the phenomenon and of the equation which expresses it. §429

According to Fourier, the 'analytical difficulty' that prevented Biot's derivation of the heat equation is caused by the absence of the concept of a heat flux:

> M. Biot was at that time unacquainted with this expression [i.e. for the heat flux], and replaced it by a differential; that is to say, he represented by an infinitely small term a physical element whose magnitude is finite. … Now

> [the heat] that flows out across the whole surface is in no way measured by a differential quantity; therefore the factor that must multiply $\mathrm{d}t$ in the expression of the transmitted heat has a finite value. One deduces this expression rigorously from the single principle proposed by Newton and Lambert.

Fourier n.d.(b), fol.157 [3]

In other words, Biot's error consisted in taking the quantity communicated at the at an interior surface to be a multiple of the differential $\mathrm{d}u = u' - u$ (rather than the gradient $\mathrm{d}u/\mathrm{d}x$), which itself comes from a careless application of Newton's law of cooling to two contiguous slices of temperature u' and u (rather than considering the heat flux at the surface). Instead, the Laplacians proposed the 'molecular hypothesis' as a solution to the analytical difficulty. Fourier does not think this solves the problem:

> The same author maintains that the difficulty which halted him and prevented him from forming a homogeneous equation cannot be removed except by supposing that heat radiates within the interior of solids. We see, on the contrary, that this difficulty disappears entirely as soon as one compares the state of a slice of the prism with that of a solid contained between two parallel and unequally heated masses [i.e. with the steady state in which the temperature varies linearly]. And in order to find the exact expression of it, one needs no physical consideration whatever about the manner in which heat propagates; it is enough to see that a solid preserves its state only when it receives as much heat as it loses.

Fourier n.d.(b), fol.157 [4]

Moreover, even setting aside whether or not the molecular hypothesis solves the analytical difficulty, Fourier is unimpressed by it and '[does] not understand why one would wish to set it up as a new truth'. The 'molecular hypothesis' is proposed as an alternative to the view that heat transfer takes place at the contact of interior surfaces in a body, which is implicit in Biot's mistaken application of Newton's law of cooling to continuous media. But, according to Fourier, this alternative view makes no sense:

Fourier 1810b in Fourier 1980 p. 37.

> We do not know that anyone has been able to imagine the movement of heat as being produced in the interior of bodies by the simple contact of the surfaces which separate the different parts. For ourselves such a proposition would appear to be void of all intelligible meaning. A surface of contact cannot be the subject of any physical quality; it is neither heated, nor coloured, nor heavy. It is evident that when one part of a body gives its heat to another there are an infinity of material points of the first which act on an infinity of points of the second. … The flow then results from an infinite multitude of actions whose effects are added.

§429. See also Fourier 1810b in Fourier 1980.

According to Fourier, the 'molecular hypothesis' does not solve the 'analytical difficulty'. Additionally, insofar as it is a physical hypothesis about the existence of molecules, it is unnecessary, and insofar as it proposes an alternative to a false view, it is trivial.

The heat flux plays no significant role in the derivation by Laplace and Poisson. The former does not consider it and the latter derives it only after deducing the heat equation. This likely is part of the reason why Fourier puts so much emphasis on its derivation and why he discusses it extensively in the summary at the end of the book.

Clifford Truesdell, in his *Tragicomical History of Thermodynamics*, is aware of the error of naively applying Newton's law of cooling to continuous media, writing: 'What Laplace disclosed to Biot, if we may accept Biot's statement, was *a new concept of heat transfer:* The conduction of heat arises not in response to differences of temperature at an actual dividing surface but to *gradients* of temperature within an undivided body'. Truesdell considers these opposing views on heat transfer to correspond to, respectively, Newton's law of cooling and Fourier's law for the heat flux. However, neither Laplace nor Fourier nor Poisson seem to have understood Newton's law of cooling in this (mistaken) way. They applied Newton's law to *molecules* which compose continuous bodies and the description of the motion of heat in terms of gradients is *derived* from this molecular perspective, by Fourier as well as Poisson. Truesdell's evidence for the assertion that Biot must have learned this new perspective on heat transfer from Laplace consists of Biot's statement of Laplace's idea: 'a particular point is influenced not only by those that touch it but also by those that are only a small distance away from it, ahead and behind' and the fact that Laplace was familiar with Euler's hydrodynamics 'which represents the accelerating force in a fluid not as a pressure difference effected by a piston but as the result of a pressure gradient within an undivided, homogeneous mass of fluid.' But Truesdell's conclusion is rather dubious in light of Laplace's explicit statements about molecules and the application of Newton's law to them, and the absence of flux-based reasoning in Laplace's (and Poisson's) derivation of the heat equation. By attributing the flux perspective to the Laplacians, he does not consider Fourier's insistence on the importance of the concept of a heat flux to be novel.

Truesdell 1980, p.50

'Molecule' in the mathematical sense, an infinitesimal part of a body.

Cited in Truesdell 1980, p.50.

Truesdell 1980, p.50

See Chapter 1 subsection III.1 and p. 24 in particular for Laplace's statement on the use of molecules.

(3) In Chapter 1 subsection III.1 we discussed how Poisson introduced a third mesoscopic scale in the derivation of the heat equation. It is defined in reference to the molecules: large enough that irregular molecular effects homogenize into something regular. Fourier, too, considers a third mesoscopic scale, but it is defined in reference to the macroscopic body: small enough so that temperatures are (sensibly) linear. The resulting hierarchy of length scales can be expressed as

$$\text{molecular spacing} \ll \text{Poisson's material part} \ll \text{Fourier's volume element} \ll \text{macroscopic body.}$$

The different choice of mesoscopic scale draws out a difference in their conceptions of the theory. The Laplacian ideal would be to derive, mathematically, the physical constants in the equation from properties of the molecules themselves. Fourier, by contrast, takes the constants as macroscopic quantities that his mathematical theory enables one to measure. For Fourier, temperature is first and foremost a macroscopic phenomenon and one descends to a finer scale only as far as the macroscopic problem demands. For Poisson, the macroscopic properties are what must be deduced from the molecular scale, which is the starting point.

I.3 Fourier as a Newtonian, and not a positivist

In the literature, Fourier's work is, with near universality, described as positivist. Consider two representative examples:

> Joseph Fourier['s] mathematical studies of heat conduction in solids, notably the Théorie analytique de la chaleur, published in 1822, was consistent with most of Comte's ideals.

Fox 1971

> For Fourier the laws of heat as formulated by him in mathematical terms represented one of the irreducible building stones of physics, and the subject of heat itself was a separate branch of physics which could not be taken over by dynamics; just as optics and electricity and magnetism represented separate branches of physics which were also incapable of explanation and absorption by dynamics. This so-called positivistic view regarding laws derived directly from observation as of primary importance and not to be explained in simpler terms was also, of course, exactly the view of Comte.

Herivel 1966. For additional examples see Chang 2004; Dhombres and Robert 1998; Grattan-Guinness 1990; Jahnke 2003. Friedman 1977 does reject the positivist identification, but rejects 'Newtonian' too, locating Fourier in the Condillacian *méthode analytique*. That is a claim about the tradition Fourier descended from; mine will be about certain philosophical fundamentals of the method.'

Such views have a long history going back to Auguste Comte himself—the creator of positivism. Comte dedicated his *Philosophie Positive* to Fourier and revered his approach to science. However, the identification of Fourier with positivism rests on a basic mistake: it packages together two fundamentally different views about the nature of scientific knowledge, which I will call *positivism* and *Newtonianism*.

Positivism is first and foremost a *negative* doctrine: its main tenet is the rejection of a certain kind of knowledge:

> In the positive state, the human mind, recognizing the impossibility of obtaining absolute notions, renounces to seek the origin and destination of the universe, and to know the intimate causes of phenomena, to devote itself solely to discovering, by the well-combined use of reasoning and observation, their effective laws, that is to say, their invariable relations of succession and similitude.

Comte 1892, p.4

Newton and the intellectual milieu of his time, too, rejected a certain kind of knowledge, the so-called 'occult qualities' championed by the Scholastics:

> The *Aristotelians* gave the Name of occult Qualities, not to manifest Qualities, but to such Qualities only as they supposed to lie hid in Bodies, and to be the unknown Causes of manifest Effects: Such as would be the Causes of Gravity, and of magnetick and electrick Attractions, and of Fermentations, if we should suppose that these Forces or Actions arose from Qualities unknown to us, and uncapable of being discovered and made manifest. Such occult Qualities put a stop to the Improvement of natural Philosophy, and therefore of late Years have been rejected. To tell us that every Species of Things is endow'd with an occult specifick Quality by which it acts and produces manifest Effects, is to tell us nothing: But to derive two or three general Principles of Motion from Phænomena, and afterwards to tell us how the Properties and Actions of all corporeal Things follow from those manifest Principles, would be a very great step in Philosophy, though the Causes of those Principles were not yet discover'd: And therefore I scruple not to propose the Principles of Motion above-mention'd, they being of very general Extent, and leave their Causes to be found out.

Newton 1952, Q.31

The rejection of occult qualities is certainly part of what Comte had in mind with positivism. But in rejecting immeasurable occult qualities, Comte considers himself to be throwing out the idea of knowledge of (fundamental) causes *as such*:

> As to determining what this attraction and gravity are in themselves, what are the causes of them, these are questions which we all regard as insoluble, which are no longer in the domain of positive philosophy, and which we rightly leave to the imagination of theologians, or to the subtleties of metaphysicians. … No righteous mind today seeks to go further.

Comte 1892, p.13. See Eșanu 2019 for Comte's more general rejection of the notion of causality as such.

Newton, unlike Comte, held that knowledge of fundamental causes—including that of gravity—is possible. Historian I.B. Cohen describes Newton's view thus:

> Newton was not a positivist; he did not say 'satis est' in the sense that he had no concern for finding causes. But he did not believe that the search for a cause, or for a mode of producing effects, should cause a halt in the application of a useful concept.

Cohen 1999, p.279

With the three-fold distinction of Scholasticism, positivism and Newtonianism in place, I will now argue that Fourier is not a positivist. Afterwards, having elaborated what Newtonianism is, I will argue that Fourier is a Newtonian.

The opening sentence of the *Théorie* reads:

> Primary causes are unknown to us; but are subject to simple and constant laws, which may be discovered by observation, the study of them being the object of natural philosophy.

This sentence is the prime exhibit for Fourier's alleged positivism. However, the immediate question it raises is: does Fourier consider the primary cause of heat (and gravity) to be merely *unknown* or—as the positivists would have it—inherently *unknowable*?

e.g. Herivel 1975, p.223, Fox 1971, p.262, Chang 2004, p.96 present it as the first line of evidence.

Before turning to this question, let us remark that the concept of a 'primary cause' itself creates ambiguity. 'Primary' is an ordinal concept; it entails that something is the first and, by definition, that nothing precedes it. Knowing that something is the primary cause of another thing means knowing that no deeper cause could be found. God and Platonic or substantial Forms are examples of things which would fit the label. However, the concept of a primary cause *as such* is incompatible with a Newtonian *inductive* approach to knowledge. The first sentence of Newton's *Principia* has a message remarkably similar to Fourier's opening sentence: 'the moderns—rejecting substantial forms and occult qualities—have undertaken to reduce the phenomena of nature to mathematical laws.' Newton's rejection of these does not imply that he considers causal knowledge, e.g. that of gravity, to be impossible. If knowledge of causes is possible and if scientific knowledge is essentially inductive, then science consists of interrelating more and more experimental evidence to induce deeper and deeper causal principles, with the most fundamental principle changing as one's context of knowledge grows. On such a view, a 'primary cause' could only be grasped at the end, i.e. if one were omniscient. On an inductive view of knowledge, it makes more sense to describe a hierarchy of causal principles with non-ordinal concepts like

Fourier would have certainly been aware that he was echoing the *Principia* preface, later in the Preliminary Discourse he even quotes its famous line: 'geometry can boast that with so few principles obtained from other fields, it can do so much'.

'*fundamental* cause', 'fundamental' being relative to a certain context of knowledge. As long as new scientific discoveries keep raising new questions in the process of answering others, there will remain unknown (but not unknowable) causes. Newton's choice of words creates less ambiguity than Fourier's; but Fourier's words—without additional context—only entail a rejection of Scholasticism; they leave ambiguous whether he is essentially Newtonian or positivist. Let us therefore evaluate Fourier's ideas, presented in subsections I.1 and I.2, against a positivist ideal.

An orthodox positivist would be one who *rejects* a molecular perspective involving unobservable entities and takes Fourier's law as the irreducible phenomenologically given starting point. Fourier is aware of the fact that one *can* develop the theory this way:

Jahnke 2003, p.199 presents Fourier's mathematical theory as starting with Fourier's law.

> I can assure you that I have often employed [molecular] considerations in my researches. But I have recognized very clearly that it was not necessary for founding the theory of heat. Everything can be reduced to a proposition for which it is easy to give a rigorous demonstration: if a solid is contained between two infinite parallel planes whose distance is e: if the temperatures of each section decrease in arithmetic progression from the interior surface up to the opposite surface, the state of the system will be permanent, that is to say it will subsist in itself once it has been set up, and there will thereafter be no change in the temperature provided that one holds the two extreme sections (A) and (B) in the states which have just been assigned to them.

See Fourier 1810b in Fourier 1980

What he proposes here as the fundamental proposition is simply Fourier's law in the case of linear temperature, from which the general result follows through mathematical deduction alone. And yet, in the *Théorie* he does *not* start here; he starts with the PCH instead, i.e. with 'molecular considerations', with a principle about the effect that distinct (infinitesimal) parts of a body have on each other. The fact that he starts with 'molecular considerations' is a conscious *choice*; he is aware that it is unnecessary. This awareness is related to the centrality of the heat flux in his theory, but this centrality does not imply a rejection of molecular considerations.

See subsection I.2.

However, the molecules of the PCH are just *mathematical* molecules, not the *physical* molecules which Laplace hypothesized. Someone who starts purely from the PCH and who considers the postulation of *physical* molecules to be unscientific might still be a proto-positivist. But the *Théorie* itself contains a discussion of the 'mode of action' of heat, which is a physical perspective pertaining to *physical* molecules: 'the molecules which compose [solid] bodies are separated by spaces void of air, and have the property of receiving, accumulating and emitting heat. Each of them sends out rays on all sides, and at the same time receives other rays from the molecules which surround it.' His criticism of the Laplacian derivation of the heat equation is that the particular 'analytical difficulty' they ran into is *not* resolved by hypothesising a certain nature of heat; instead, it is resolved by the concept of a heat flux. Additionally, we saw that Fourier, in his private notes, initially considered the nature and mode of action of heat as a single unit, but then separated them and (falsely) concluded that the universal mode of action of heat is radiation,

See subsection I.1.

§44

without speculating about its nature. Most crucially, Fourier *does* have a perspective on the nature of heat: its nature is like that of light. To appreciate the theoretical difficulty Fourier was faced with, we have to keep in mind that Fourier considered 'heat' to be a single kind of phenomenon—radiant heat and conducted heat are both forms of *heat*. Consider then the developments in physics required for a unified theory of heat; one would have,

(1) for conduction: to develop thermodynamics and the kinetic theory: energy conservation, entropy, the mechanical equivalent of heat, heat in matter as molecular motion;
(2) for radiation, completely separate from that: to develop electromagnetism and the field concept: conceiving radiation as electromagnetic waves (rather than caloric emission);
(3) to develop statistical mechanics: a mechanical-probabilistic account of entropy and the second law that generalizes the kinetic theory;
(4) to apply (2) and (3) together to thermal radiation, which shows that radiant heat, too, is subsumed under the unified theory.

Fourier, of course, couldn't have been aware of the need for such theoretical advances. But, with the benefit of hindsight, we can see that an integrated theoretical account of heat required more ideas than can be expected of any one man. Moreover, we can appreciate that either Fourier had to accept that heat was in fact two distinct things, something that seemed to contradict observations he considered basic (e.g. the tendency toward equilibrium is common to both) or he would have to set aside what he knew about radiant heat—precisely the ideas on which he was drawing for his physical understanding of heat—to adopt the kinetic or caloric theory. Fourier didn't have a comprehensive physical theory of heat, but his refusal to speculate comes out of his judgement that the evidence for either is *insufficient*, not that the investigation is *unscientific*. If Fourier's refusal to speculate about the nature of heat made him a positivist then 'positivist' becomes a term for people with fundamentally different views: those who hold that the nature of heat is unknowable and those who consider it to be unknown (but knowable). The perspective of Fourier as a positivist does not consider Newtonianism a fundamental third alternative to Scholasticism and positivism.

So far I have argued that both Newton and Fourier hold that knowledge of causes is possible and that they therefore should not be considered positivist. But Newtonianism is a particular view about a particular way by which knowledge of causes can be derived from phenomena (using mathematics). It is this which distinguishes Newtonianism from Scholasticism. To assess whether Newtonianism is a genuine third alternative and to asses whether Fourier is a Newtonian, we have to consider Newton's method in more detail. To this end, I will draw on George Smith's celebrated account of Newton's method in the *Principia* and argue that it also describes Fourier's method in the *Théorie*. Smith 2002

Smith contrasts Newton's method against the hypothetico-deductivism of Newton's contemporaries such as Descartes and Huygens, for whom a theory was confirmed by deducing observable consequences from conjectured hypotheses and checking them against the phenomena. This contrast, of course, follows Newton's own self-positioning against the 'mechanical philosophers', specifically his refusal to 'feign hypotheses' and his insistence that propositions be deduced from the phenomena. Both were presenting themselves as genuine alternatives to the Scholastics, who were universally rejected. Newton's views were not universally accepted:

> To his contemporaries, what seemed most confusing about Newton's way of talking about forces was his willingness to put forward a theory of gravitational 'attraction' without regard to the causal mechanism effecting it. They generally concluded that he had to be committed to action at a distance as a causal mechanism in its own right. The outspoken opposition to the *Principia* in many quarters stemmed primarily from the inexplicability of action at a distance.

Smith 2002, p.203

The criticism is that, because Newton did not specify a mechanical contact mechanism for gravity, it is 'beyond explanation and hence occult'. This would make Newton no better than the Scholastics he claims to reject. But if we reject such criticism, if we accept that we can know that gravity exists without knowing the mechanism by which it operates, then that still does not tell us how Newton gains knowledge of it. How *do* we gain knowledge of it? Basic to Newton's method is a fundamental distinction between considering phenomena *mathematically* and considering them *physically*. Continuing Smith's quote:

Smith 2002, p.202

> Present-day readers, viewing the *Principia* in the light of 300 years of success in physics, are not likely to find the way Newton talks of forces from a physical point of view confusing. What most tends to confuse them is the distinction between considering forces from a physical point of view and considering them purely from a mathematical point of view.

Considering 'attraction' and 'impulse' purely mathematically means considering 'not the species of forces and their physical qualities but their quantities and mathematical proportions'. Such a perspective itself is founded on basic empirical observations about motions (e.g. that there is such a thing as friction, which one needs to know to formulate the laws of motion), but not on a sophisticated physical *theory*. The object of Fourier's *Théorie* is 'the problem of the propagation of heat', which 'consists in determining what is the temperature at each point of a body at a given instant, supposing that the initial temperatures are known.' But what is the purpose of such a theory? More specifically, why distinguish a mathematical and physical perspective on one single phenomenon? Quoting from the Scholium of the *Principia* that Smith designates as Newton's description of his distinctive method:

Newton 1999, Bk. 1, Sec. 11

§1

> Mathematics requires an investigation of those quantities of forces and their proportions that follow from any conditions that may be supposed. Then, coming down to physics, these proportions must be compared with the phenomena, so that it may be found out which conditions of forces apply to each kind of attracting bodies. And then, finally, it will be possible to argue more securely concerning the physical species, physical causes, and physical proportions of these forces.

Newton 1999, Bk. 1, Sec. 11

The mathematical theory is thus more general—since it considers forces *qua* their 'quantities and mathematical proportions'—yet it is also *prior* to the physical theory, which identifies the kind of forces that exists in nature. The physical theory is developed by *applying* the mathematical theory. The mathematical theory affords what Smith has called *theory-mediated measurement*; such measurement then allows us to argue 'more securely' about physical phenomena. The aim of the 'mathematical theories of motion developed in Books 1 and 2 of the *Principia*'

> is to provide a basis for specifying experiments and observations by means of which the empirical world can provide answers to questions—this in contrast to conjecturing answers and then testing the implications of these conjectures. Newton is using mathematical theory in an effort to turn otherwise recalcitrant questions into empirically tractable questions.

Smith 2002, p.197–8

This is also what Fourier considers himself to be doing in the *Théorie*. In the *Preliminary Discourse* he describes his theory as moving from observation to a small set of simple facts, and from those to measurement:

> To found the theory, it was in the first place necessary to distinguish and define with precision the elementary properties which determine the action of heat. I then perceived that all the phenomena which depend on this action resolve themselves into a very small number of general and simple facts; whereby every physical problem of this kind is brought back to an investigation of mathematical analysis. From these general facts I have concluded that to determine numerically the most varied movements of heat, it is sufficient to submit each substance to three fundamental observations. Different bodies in fact do not possess in the same degree the power to *contain* heat, to *receive* or *transmit it across their surfaces*, nor to *conduct* it through the interior of their masses. These are the three specific qualities which our theory clearly distinguishes and shews how to measure.

Preliminary Discourse

The theory thus affords the *measurement* of these physical qualities. In the concluding articles of the *Théorie*, he names 'the differential equations of the movement of heat in solid or liquid bodies, and the general equation which relates to the surface' as the 'chief results' of his theory, writing that

> the truth of these equations is not founded on any physical explanation of the effects of heat. In whatever manner we please to imagine the nature of this element, whether we regard it as a distinct material thing which passes from one part of space to another, or whether we make heat consist simply in the transfer of motion, we shall always arrive at the same equations, since the hypothesis which we form must represent the general and simple facts from which the mathematical laws are derived.

§432

The mathematical theory is thus *not* founded on a physical theory of heat, it is derived from 'general and simple facts'. Their relationship is the other way around: the physical theory has to be formed in such a way as to represent the results established by the mathematical theory. At the end of the *Preliminary Discourse*, he speaks directly to what lies beyond his work:

> The new theories explained in our work are united for ever to the mathematical sciences, and rest like them on invariable foundations; all the elements which they at present possess they will preserve, and will continually acquire

> greater extent. Instruments will be perfected and experiments multiplied. The analysis which we have formed will be deduced from more general, that is to say, more simple and more fertile methods common to many classes of phenomena. For all substances, solid or liquid, for vapours and permanent gases, determinations will be made of all the specific qualities relating to heat, and of the variations of the coefficients which express them. At different stations on the earth observations will be made, of the temperatures of the ground at different depths, of the intensity of the solar heat and its effects, constant or variable, in the atmosphere, in the ocean and in lakes; and the constant temperature of the heavens proper to the planetary regions will become known. The theory itself will direct all these measures, and assign their precision.

Preliminary Discourse

What Fourier is envisioning is that his theory *directs* the improvement of measurement equipment, the measurement of the 'specific qualities relating to heat' for 'all substances', and the measurement of temperatures of the earth, the solar heat, the oceans and lakes—he is describing how his work affords theory-mediated measurement of physical phenomena. Moreover, such measurement also allows the theory to develop further in what Smith has described as the use of 'systematic deviations from current theory as evidence in a process of successive approximations'. Having first drawn out the consequences of the theory under simplifying assumptions, Fourier holds that 'careful comparison of the results with those of very exact experiments' will reveal 'what corrections must be employed', and that it is in this way that we shall 'ascertain what are the causes which modify the movement of heat'. Writing to the Swiss physicist Pierre Prévost, Fourier makes the same point about gravity. Prévost had proposed, in his letter, that gravity originates in a radiant matter. Fourier grants this possibility while noting that the mathematical theory already suffices for measurement:

Smith 2002, p.212

§433

> It may well be that the origin of this force is due to a radiant matter; but here all the experiments have been made, and the whole mathematical theory is formed. Knowledge of the cause has been replaced by its mathematical expression, which almost always suffices for the measurement of phenomena. You will bring minds back to considerations of a different order, and it will have been given to you to contend with two great powers [grandes puissances]: gravity and heat.

The 1817 letter is found in Prévost 1832, p.93.

The expression suffices for measurement 'almost always', presumably the exception being where the 'causes which modify the movement of heat' re-enter. Moreover, investigation into physical causes is not dismissed but assigned to 'considerations of a different order'.

I thus consider Fourier to be a Newtonian. Newtonianism, so understood, ought to be considered a genuine third alternative to Scholasticism and positivism because it holds both (a) that that mathematical theories of natural phenomena can be founded solely on what is empirically given (agreeing with positivism), and (b) that such theories can be the foundation for physical theories that give deep causal knowledge of the phenomena (disagreeing with positivism). In naming a doctrine at this level of abstraction I am deliberately not trying to account for all beliefs which Newton held. There may be views of his—say, his views on absolute space and time—that are (or are not) in tension with the method I am isolating, I am also not claiming a comprehensive account of his method; doing either would require an analysis not possible in this thesis, as would a more thorough treatment of Scholasticism and positivism as historical movements. Nevertheless, speaking at this level of generality reveals a single method shared by Fourier and Newton—a mathematical theory founded on empirically given facts, formed for the sake of measuring the phenomena to advance toward knowledge of their causes—a method that distinguishes both of them, to various degrees, from their contemporaries. Fourier was contrasting his views with the Laplacians' mathematical theory of heat. They sought to found it on a physical hypothesis—the caloric theory—which itself is based on Laplace's broader hypothesis that 'the phenomena of nature are in the last analysis reduced to actions *ad distans* from molecule to molecule'. The Laplacians are therefore Newton's heirs in content: they inherit the particular kind of theoretical explanation—in terms of particles interacting by means of forces—which Newton advocated; yet, in method, they are using the exact hypothetico-deductivism which Newton repudiated. Fourier, on the other hand, is Newton's heir in method: he, too, distinguished a mathematical and physical perspective on natural phenomena and—having founded the mathematical theory on basic observations—sought to base the physical investigation of heat on the theory-mediated measurement that the mathematical theory affords.

See Chapter 1 subsection III.1.

Newton's use of the mathematical theory of Books 1 and 2 of the *Principia* to answer particular physical questions *also* leads him to a fundamental *physical* insight: the universal law of gravitation. Does Fourier's mathematical theory of heat also reveal some fundamental physical perspective on heat? It does not distinguish between the caloric and kinetic theories of heat. More broadly, it does not reveal any microscopic characteristics of heat. However, it does reveal a *macroscopic* one: the 'simple modes' in which heat moves. Recalling remark (1) of Chapter 1 section I, Fourier considers the decomposition of a temperature distribution into its 'simple modes' to be more than a mere mathematical decomposition (e.g. the decomposition of a force into orthogonal directions). The decomposition describes something physically real: 'the natural effect whose laws we seek, is really decomposed into distinct parts, corresponding to the different terms of the series' and the 'essential character' of Fourier's solution is the 'representation of phenomena'. 'In this composition we ought not to see a result of analysis due to the linear form of the differential equations, but an actual effect which becomes sensible in experiments.' However, a discussion of Fourier's physical perspective on these modes first requires a more substantial account of his mathematical theory, to which we now turn.

§428

§428

His physical perspective on the modes is taken up in subsection II.4.

II The form of the solution to the heat equation

The Laplacians criticized Fourier's derivation of the heat equation, but there was agreement about the resulting equations, the PDE as well as the BC. This conveniently separates their disagreements into two parts.

We will now turn to the second part and consider Fourier's method for solving the heat equation, comparing it to Poisson's method discussed in Chapter 1 subsection III.2.

The difference between Fourier and Poisson starts, not with the answer, but with the nature of the problem itself. We see, in Fourier, the birth of the *initial boundary value problem* (IBVP): to find a function satisfying the PDE, boundary condition *and* initial condition, not just a function that satisfies a PDE. Fourier concludes his derivation of the problem for the sphere as follows:

> the three equations which are required to determine the function $\phi(x,t)$ or v are the following,
>
> $$\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{K}{CD}\left(\frac{\mathrm{d}^2 v}{\mathrm{d}x^2} + \frac{2}{x}\frac{\mathrm{d}v}{\mathrm{d}x}\right), \qquad K\frac{\mathrm{d}v}{\mathrm{d}x} + hv = 0, \qquad \phi(x,0) = 1.$$
>
> The first applies to all possible values of x and t; the second is satisfied when $x = X$, whatever be the value of t; and the third is satisfied when $t = 0$, whatever be the value of x. §117

The three equations form a single unit defining a single problem. Poisson, by contrast, essentially solves two problems: first, to determine the integral of the PDE (i.e. the so-called fundamental solution); subsequently, to apply the integral to more particular problems involving special conditions.

Fourier does consider the heat equation on an infinite domain where no boundary conditions apply. But, fundamentally, he considers this one particular (though special) problem—the case of a ring of infinite radius—rather than a different more general problem. See §397–8, §428 n.10.

II.1 The method of separation of variables

Fourier does not start by defining the problem for an arbitrary geometric domain and then applying the general form to particular geometries, since

> this method would often involve very complicated calculations which may easily be avoided. There are several problems which it is preferable to treat in a special manner by expressing the conditions which are appropriate to them. §101

He derives general equations for the problem—ones with a boundary of arbitrary shape in terms of Cartesian coordinates—only after first considering particular cases. To subsequently solve the problem, he uses in Chapter 3 a model problem of the steady-state solution half-strip heated at one end to 'indicate the elements of the method which we have followed'; Chapter 4 then considers the ring; Chapter 5 the sphere; Chapter 6 the cylinder; Chapter 7 the half-prism; Chapter 8 the cube; and lastly, Chapter 9 the whole space, i.e. the case without boundary conditions (with applications of similar methods to other PDEs).

A half-strip is defined by: $|y| < \pi/2$, $z > 0$.

The half-prism is the half-strip with an additional dimension $|x| < \pi/2$.

In each case, the solution is obtained by the same method, today known as separation of variables. 'Separation of variables' is not a very good name for this method however, since it names the method after a minor step of a general method for solving IBVPs from start to finish. A better name would be 'the method of simple modes' (or, permitting technical jargon, 'the spectral method'); it is by reference to 'simple functions' that Fourier first introcues the method:

> In order to consider the problem in its elements; we shall in the first place seek for the simplest functions of x and y, which satisfy equation (a); we shall then generalise the value of v in order to satisfy all the stated conditions. By this method the solution will receive all possible extension, and we shall prove that the problem proposed admits of no other solution.

(a): the relevant PDE, in this case the simplified equation $\frac{\mathrm{d}^2 v}{\mathrm{d}x^2} + \frac{\mathrm{d}^2 v}{\mathrm{d}y^2} = 0$.

In the model problem of the half-strip he explicitly separates the variables, writing $v = F(x)f(y)$ and deriving two separate ODEs linked by a constant, but in subsequent problems he will jump straight to the ansatz $v = e^{mt}u$. To substantiate the method, let us consider the problem of the sphere.

Sphere problem: $\begin{cases} \frac{\mathrm{d}v}{\mathrm{d}t} = k(\frac{\mathrm{d}^2 v}{\mathrm{d}x^2} + \frac{2}{x}\frac{\mathrm{d}v}{\mathrm{d}x}), \\ \frac{\mathrm{d}v}{\mathrm{d}t} + hv = 0, \\ v = F. \end{cases}$ where the second applies only to $x = X$ and the third to $t = 0$.

First, Fourier finds the simple solutions

$$v = \frac{e^{-kn^2 t}}{x}(A\cos(nx) + B\sin(nx)).$$

§284

for different n. He concludes that $A = 0$ since otherwise the temperature is infinite at the origin. These solutions are simple since each evolves 'separately as if it alone existed'; if such a simple state were formed 'it would subsist of itself, and this is the property which distinguishes it from all other states'.

§203

§164

Subsequently, by substituting the simple solution in the BC, he derives a condition on n:

$$\frac{nX}{\tan(nX)} = 1 - hX.$$

He finds that the equation has an infinite number of real roots corresponding to particular values of nX, where

> the first is included between 0 and $\frac{\pi}{2}$, the second between π and $\frac{3\pi}{2}$, the third between 2π and $\frac{5\pi}{2}$, and so on. These roots approach very near to their upper limits when they are of a very advanced order. §285

Denoting the obtained roots by $n_1, n_2, n_3 \ldots$, he then superposes the found simple solutions to derive a more general solution

$$v \cdot x = a_1 e^{-kn_1^2 t}\sin(n_1 x) + a_2 e^{-kn_2^2 t}\sin(n_2 x) + a_3 e^{-kn_3^2 t}\sin(n_3 x) + \ldots. \quad (1)$$

Lastly, 'it remains only to form any initial state by means of a certain number, or of an infinite number of partial states.' By taking $t = 0$ in Eq. (1), this comes down to defining $a_1, a_2, a_3, \ldots$ such that §290

$$F(x) = \frac{1}{x}\Big(a_1\sin(n_1 x) + a_2\sin(n_2 x) + a_3\sin(n_3 x) + a_4\sin(n_4 x)\ldots\Big) \quad (2)$$

He finds that

$$a_i = \frac{1}{X - \frac{1}{2n_i}\sin(2n_i X)} \int_0^X x\sin(n_i x)F(x)\,\mathrm{d}x.$$

$\int_0^X \sin^2(n_i x)\,\mathrm{d}x = X - \frac{1}{2n_i}\sin(2n_i X)$

This solves the problem if we assume that an arbitrary initial state $F(x)$ can be decomposed into a series like Eq. (2). Filling the a_i into Eq. (1), Fourier concludes that

> this is the form which must be given to the general integral of the equation $\frac{\mathrm{d}v}{\mathrm{d}t} = k(\frac{\mathrm{d}^2 v}{\mathrm{d}x^2} + \frac{2}{x}\frac{\mathrm{d}v}{\mathrm{d}x})$ in order that it may represent the movement of heat in a solid sphere. §291

This is followed by a uniqueness proof (in Chapter 3 only) from which

> it follows that the problem proposed admits no other solution than that which results from equation (α). Another form may be given to this result, but the solution can be neither extended nor restricted without rendering it inexact.

§202–4

This is the essence of Fourier's method, consisting of three steps: (1) find the simple solutions (eigenfunctions); (2) use the BC to define the constants of the simple solutions (eigenvalues); and (3) decompose the initial state into a series of functions defined by the first two steps.

Until now I have described *what* Fourier claimed, but I have left aside what kind of mathematical justification he provided for these claims. We will turn to the mathematical aspects in subsections II.2 and II.3 and subsequently its physical aspects in subsection II.4, where we contrast Fourier with Poisson and discuss why Fourier considered his form of the solution to be superior to others.

II.2 Mathematical aspects of Fourier's solution

See Grattan-Guinness 1990, p.148.

In 1755 Daniel Bernoulli had already claimed that the most general solution to the wave equation $\frac{d^2y}{dt^2} = \frac{d^2y}{dx^2}$ subject to the BC: $y(\pm\pi) = 0$ is the series $\sum_{n=1}^{\infty} a_n \sin(nt)\sin(nx)$. By taking $t = 0$, this claim logically depended on the decomposition of the (arbitrary) initial state into a series of trigonometric functions. However, the only kind of justification which Bernoulli's claim rested on is the fact that the solution has infinitely many degrees of freedom. Presumably, these could be used to satisfy the infinite degrees of freedom of the initial state. One of Fourier's key contributions consisted in actually defining the coefficients and explaining what their meaning was, though Euler had already identified the coefficient formula, something initially unbeknownst to Fourier. Commenting on Bernoulli's generic claim, Fourier remarked that

> the most complete of all the proofs of this proposition is that which consists in actually resolving a given function into such a series with determined coefficients.

§230

More significantly, it is Fourier who truly turns the idea of describing a solution in the form of a series of 'simple modes' into a general and practical method for solving a broad class of mathematical problems, and it is Fourier who recognizes the mathematical properties which form the foundation of this method: the reality of the roots of the transcendental equation that defines the frequencies in the case of so-called convective boundary conditions, the ability to decompose an arbitrary function into a series of 'simple modes', the orthogonality of these simple modes and the need to prove uniqueness of the solution.

Within the solution procedure as a whole, the proof of two mathematical propositions had been controversial and these propositions are given significant attention in the *Théorie*: that all roots of the transcendental equation are real and that a series decomposition is possible for arbitrary functions. We will not discuss Fourier's (various) proofs of the former and, given the scope of the latter, it will be considered in section III.

Fourier's arguments for the reality of the roots are found in §§285–8 and §305

Fourier also cleared various conceptual roadblocks that had made others skeptical of the generality of his series. These conceptual issues will also be taken up in section III.

By considering the heat equation in different geometries and by considering convective boundary conditions, he went beyond mere decomposition of a function into a series of trigonometric functions of integer frequency. Each geometric shape corresponds to a different solution, but they are solved by the same method and the different solutions have a common mathematical form:

$$u(x,t) = \sum_{n=1}^{\infty} a_n e^{-\mu_n^2 t} \phi(\mu_n x) \tag{3}$$

where $\phi(\mu_n x)$ is $\frac{\sin(\mu_n x)}{\mu_n x}$ in the case of the sphere, $A\cos(\mu_n x) + B\sin(\mu_n x)$ in the case of the cube, and $J_0(\mu_n x)$—what are today called Bessel functions of the first kind—in the case of the cylinder. The μ_n are determined by the BC and the a_n by the IC. Fourier is aware of the generality and uniformity of his method:

In case of the sphere and cylinder we don't consider angular temperature variations. I restricted the cube to one dimension to make the contrast more clear.

> If we compare with each other the solutions relative to the varied movement of heat in a ring, a sphere, a rectangular prism, a cylinder, we see that we had to develop an arbitrary function $f(x)$ in a series of terms, such as
>
> $$a_1\phi(\mu_1 x) + a_2\phi(\mu_2 x) + a_3\phi(\mu_3 x) + \ldots$$
>
> The function ϕ, which in the second member of equation (A) is a cosine or a sine, is replaced here by a function which may be very different from a sine. The numbers $\mu_1, \mu_2, \mu_3 \ldots$ instead of being integers, are given by a transcendental equation, all of whose roots infinite in number are real.

(A): $f(x) = \frac{1}{2\pi}\sum\int_a^b f(\alpha) \cos\left(\frac{2i\pi}{X}(x-\alpha)\right)\mathrm{d}\alpha$.

§424

This leads us to what I am calling Fourier's *spectral perspective* on his solutions.

★ II.3 Fourier's spectral perspective

To shed a light on the significance of Fourier's contribution to mathematics from a deeper theoretical perspective, let us consider its modern descendant—the spectral theorem.

Theorem 2.1 (Spectral theorem for compact self-adjoint operators)**.** Let H be a separable Hilbert space and $A : H \to H$ a compact self-adjoint operator. Then:

For a proof see Conway 2007.

(1) The spectrum $\sigma(A)$ consists of real eigenvalues, at most countably many, with 0 as the only possible accumulation point. Each nonzero eigenvalue has finite-dimensional eigenspace.

(2) There exists an orthonormal basis $\{e_n\}_{n\in\mathbb{N}}$ of H consisting of eigenvectors of A, with $Ae_n = \lambda_n e_n$, so that

$$Ax = \sum_n \lambda_n \langle x, e_n\rangle e_n \qquad \text{for all } x \in H,$$

where the series converges in H and $\lambda_n \to 0$.

Corollary 2.1.1 (Spectral decomposition for linear evolution equations)**.** Let L be a self-adjoint operator on a separable Hilbert space H, bounded below, with compact resolvent. Then:

e.g. $L = -\Delta$ on a bounded domain $\Omega \subset \mathbb{R}^d$ with convective boundary conditions; compactness of the resolvent follows from elliptic regularity and Rellich–Kondrachov.

(1) The spectrum $\sigma(L)$ consists of real eigenvalues, at most countably many, with $+\infty$ as the only possible accumulation point. Each eigenvalue has finite-dimensional eigenspace.
(2) There exists an orthonormal basis $\{e_n\}_{n\in\mathbb{N}}$ of H consisting of eigenvectors of L, with $Le_n = \lambda_n e_n$, so that for any $u_0 \in H$ the evolution equation

$$\partial_t u = -Lu, \qquad u(0) = u_0,$$

has the solution

$$u(t) = \sum_n e^{-\lambda_n t} \langle u_0, e_n \rangle \, e_n. \tag{4}$$

Fourier's work obviously does not have the generality of the spectral theorem; but more fundamentally, Fourier lacked the concepts necessary to conceive of such a general statement. Essentially all concepts entering into these statements—separable Hilbert space, compact self-adjoint operators, spectrum, finite dimensional eigenspace, etc.—would have been foreign to Fourier. The similarity between the spectral theorem and Fourier's work is structural: Fourier considered various particular IBVPs defined for arbitrary functions, he argued (1) that the roots of the associated transcendental equation are real, discrete and tend to infinity and (2) that the class of 'simple modes' which he found in each case are orthogonal, that they can be used to decompose any arbitrary function and, subsequently, that the IBVPs have a solution of the form Eq. (4). This summary of Fourier's work has essentially the same structure as Corollary 2.1.1, the difference is in content: the concepts which enter into Corollary 2.1.1 are both abstract and general as against Fourier's particular instances (self-adjoint operator instead of Fourier's three PDEs), and they formalize ideas which Fourier treated more informally ('arbitrary function' versus element of a Hilbert space).

We will use four key developments post-Fourier to draw out mathematical aspects of Fourier's method which deserve more commentary: (1) Sturm-Liouville theory, which defined the general class of IBVPs to which Fourier's method applies; (2) Green's functions, which invert the PDE into integral equations, making the mathematics more tractable; (3) vector calculus and coordinate-free multivariable calculus, which defines the relevant operators purely geometrically rather than in fixed coordinate systems; (4) functional analysis: infinite dimensional linear algebra, which defines the Hilbert space, the class of functions to which the theorem applies. Added to these is a remark on something present in Fourier but absent in the spectral theorem: (5) the case of the unbounded domain.

(1) Fourier is aware of the fact that his method extends beyond the particular cases he has solved. However, he has no description of the general class to which his methods apply: he has not identified what property the particular IBVPs which he considered have in common in virtue of which his method applies. This was left to his protégé Sturm and compatriot Liouville. They identified a broader class of IBVPs to which Fourier's method applies: whenever, upon separating the variables, each function

Problem	(p, q, w)	Interval
Heat in bar	$(1, 0, 1)$	$(-L, +L)$
Radial heat (sphere)	$(r^2, 0, r^2)$	$(0, R)$
Radial heat (cylinder)	$(r, 0, r)$	$(0, R)$
Gravitational potential (sphere)	$(1-\mu^2, \frac{m^2}{1-\mu^2}, 1)$	$(-1, +1)$
Urn (Bernoulli–Laplace)†	$(e^{\mu^2}, -2e^{\mu^2}, e^{\mu^2})$	$(-\infty, +\infty)$

†Separating $u = e^{-\lambda r'}\phi(\mu)$ reduces the urn PDE to $\phi'' + 2\mu\phi' + (2+\lambda)\phi = 0$, which is transformed into $-(e^{\mu^2}\phi')' - 2e^{\mu^2}\phi = \lambda e^{\mu^2}\phi$.

Table 1. Particular Sturm–Liouville problems arising in Fourier's Laplace's problems.

of one variable is defined by an ODE of the form

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}u}{\mathrm{d}x}\right) + q(x)u = \lambda w(x)u, \qquad a < x < b, \tag{5}$$

subject to the BC

$$\alpha_1 u(a) + \alpha_2 p(a)u'(a) = 0, \qquad \beta_1 u(b) + \beta_2 p(b)u'(b) = 0.$$

Here $(\alpha_1, \alpha_2) \neq (0,0)$ and $(\beta_1, \beta_2) \neq (0,0)$.

The three forms of the heat equation which Fourier considered all give rise to ODEs of this form. Moreover, so do the two PDEs Laplace derived in the course of studying the gravitational potential and the urn problem! From Laplace to Fourier to Sturm and Liouville we see a genuine growth in understanding: Laplace identified the orthogonal property of his simple modes and decomposed arbitrary functions in terms of them, however he never forms a *general* concept like 'simple mode' and seems unaware of a common theme across his two decompositions. Fourier did understand that there was a common theme: he formed the concept of a 'simple mode' and stated in general form some of the relevant properties. However, he did not identify the fundamental property that the particulars have in common and of which the other properties are a consequence. Sturm and Liouville do with their ODE. The difference in their understanding is most clearly seen in the way they prove the properties. Fourier proves separately that the sines are orthogonal and that the Bessel functions are orthogonal, even though he recognizes that this orthogonality is common across the various decompositions. Sturm and Liouville on the other hand, do not use the concrete eigenfunctions, but derive directly from general ODE the orthogonality of *any* particular instance corresponding to particular coefficients and BC.

See Table 1.

Sturm and Liouville's ideas can be formulated more abstractly in terms of concepts of linear algebra. This is how they enter the spectral theorem. Dividing the LHS of Eq. (5) by w gives an operator L that is self-adjoint for the inner product $\langle f, g\rangle_w = \int_a^b f\bar{g}w\,\mathrm{d}x$, the separated BCs being exactly what makes the boundary term from integrating by parts vanish. Consequently, the reality of the eigenvalues and the orthogonality of the eigenfunctions can be derived by (rather simple) linear algebraic arguments. *Self-adjointness* is the operator-theoretic perspective on the property Sturm and Liouville had identified, and it is in this form that it enters the spectral theorem.

L is *self-adjoint* if $\langle Lu, v\rangle_w = \langle u, Lv\rangle_w$ for all u, v. It is the infinite-dimensional analogue of a symmetric (Hermitian) matrix, whence the real spectrum and orthogonal eigenvectors.

(2) Fourier and his compatriots considered the heat equation first and foremost in the form of a PDE. Read in the language of operators, the modern theory proceeds the other way around, replacing the differential operator L with its inverse integral operator—the Green's function. Theorem 2.1 applies first and foremost to this integral operator, not to a PDE.

The motivation is that the differential operator is unbounded, but the integral operator that inverts it is bounded.

Constructing an integral equation from the PDE is precisely what Poisson did in Chapter 1 subsection III.2! However, the solution operator of the modern theory pertains to an IBVP on a bounded domain: the BC are part of what define the kernel, they are not imposed on a general solution afterwards. In this respect the modern theory takes up Fourier's conception of the IBVP as forming a single unit (section II), rather than Poisson's conception of the PDE alone as forming the most general problem. Moreover, both Fourier and Poisson were after something more specific than the modern theory: concrete analytic expressions, be they integrals or series. Fourier's method determines concrete eigenfunctions in closed form, from which a solution operator can be constructed. But this is only possible for the simplest, most symmetric geometries; for a general domain neither the eigenfunctions nor the kernel of the solution operator have a closed-form expression. Distinctive to the modern theory is that one can establish that the solution operator exists and affords a spectral decomposition *without* explicit analytical formulae; the spectral data can then be approximated afterwards, e.g. by approximating the domain with the simple shapes for which they are known.

(3) Fourier considered the heat flux in the form of specific coordinate systems—rectangular for the cube, spherical for the sphere, cylindrical for the cylinder—re-deriving the relevant form of the heat equation in each by reasoning about fluxes through volume elements expressed in analytical coordinates. He gets the equations right but has to re-do the calculus three times. The mathematical framework to treat fluxes purely geometrically and coordinate-free did not exist. The development of (multivariable) vector calculus as a discipline distinct from single variable calculus, with its own characteristic objects (gradient, divergence, Laplacian) and theorems (Stokes, divergence theorem) happened later in the nineteenth century. Vector calculus is what allows the Laplacian to be defined uniformly across geometries, so that the spectral theorem applies to heat flow on the bar, the sphere, and the cylinder via a single operator rather than three coordinate-specific expressions.

The development of vector calculus strictly speaking, not necessary for the spectral theorem itself. Corollary 2.1.1 is a statement of pure functional analysis: L is a linear map on a Hilbert space. Vector calculus enters when one considers L to be a differential operator on a geometric domain.

(4) There remains the contested issue of which functions the theory applies to. Fourier considers it to apply to 'arbitrary functions', which of course raises the question of what 'arbitrary function' means. The spectral theorem, by contrast, says: elements of a Hilbert space; for PDEs on geometric and bounded domains, the set of square-integrable integrable Lebesgue functions L^2 deserve mention. In subsection III.4 we will discuss Fourier's conception of 'function' and see that the perspective of both Laplace and Fourier on 'function' is, though informal, essentially that of an infinitely long vector. This view itself comes out of the idea of treating curves as an infinitangular polygons. I will restrict myself to pointing out that the theory of Hilbert spaces is a theory of infinite-dimensional

linear algebra. This theory was developed at the start of the 20th century when figures like Fredholm, Hilbert and Schmidt extended ideas of finite-dimensional linear algebra to infinite dimensions. In Fourier's time there was not even a systematic theory of finite-dimensional linear algebra, though the practice of stating and solving linear systems of equations was common. See Jahnke 2003, Ch.13.

(5) Fourier considered the movement of heat on an infinite domain simply by taking the limit of his solution in the finite-domain case, turning a series decomposition into an integral. The modern theory applies first and foremost to *bounded* domains: on $\mathbb{R}$ the operator $-\partial_x^2$ has no eigenpairs; $e^{i\sqrt{\lambda}x}$ solves $-\partial_x^2 u = \lambda u$, yet it is not an eigenfunction in a Hilbert space, since $e^{i\sqrt{\lambda}x} \notin L^2(\mathbb{R})$. Fourier was not working with function spaces defined by a norm, hence there is nothing to rule these out; his 'simple functions' remain well-defined on an infinite domain. Attempting to emulate Fourier's move of directly taking the limit of the bounded domain would require additional theory today: if $e^{i\sqrt{\lambda}x}$ are no longer considered as functions but as distributions, then it is possible. See §397–8 for his limit argument.

II.4 Physical aspects of Fourier's solution: Geometrical shape as a cause

Fourier thought he had rigorously justified all aspects of his solution. However, independent of its mathematical justification, Fourier considered his solutions to describe something physically real:

> We have said that each of these solutions gives *the equation proper to the phenomenon*, since it represents it distinctly throughout the whole extent of its course, and serves to determine with facility all its results numerically. The functions which are obtained by these solutions are then composed of a multitude of terms, either finite or infinitely small: but the form of these expressions is in no degree arbitrary; it is determined by the physical character of the phenomenon. For this reason, when the value of the function is expressed by a series into which exponentials relative to the time enter, it is of necessity that this should be so, since the natural effect whose laws we seek, is really decomposed into distinct parts, corresponding to the different terms of the series. The parts express so many *simple modes* compatible with the special conditions; for each one of these modes, all the temperatures decrease, preserving their primitive ratios. In this composition we ought not to see a result of analysis due to the linear form of the differential equations, but an actual effect which becomes sensible in experiments. §428

We have remarked before that the mere decomposition of a mathematical quantity does not make the constituent parts physically significant. However, Fourier is arguing that the simple modes into which a temperature distribution is decomposed *are* physically significant. Each 'simple mode' behaves 'as if alone existed'. The various forms his solutions take thus have *objectively* the right form: See remark (1) in Chapter 1 section I. He uses this expression in §170, §191, §203, §290.

> The integrals which we have obtained are not only general expressions which satisfy the differential equations; they represent in the most distinct manner the natural effect which is the object of the problem. This is the chief condition which we have always had in view, and without which the results of investigation would appear to us to be only useless transformations. When

> this condition is fulfilled, the integral is, properly speaking, *the equation of the phenomenon*; it expresses clearly the character and progress of it, in the same manner as the finite equation of a line or curved surface makes known all the properties of those forms. To exhibit the solutions, we do not consider one form only of the integral; we seek to obtain directly that which is suitable to the problem. Thus it is that the integral which expresses the movement of heat in a sphere of given radius, is very different from that which expresses the movement in a cylindrical body, or even in a sphere whose radius is supposed infinite. Now each of these integrals has a definite form which cannot be replaced by another. It is necessary to make use of it, if we wish to ascertain the distribution of heat in the body in question. In general, we could not introduce any change in the form of our solutions, without making them lose their essential character, which is the representation of the phenomena. §428

Poisson in 1815, by contrast, had insisted on using one particular form, the fundamental solution

In this subsection I will refer to Poisson's 1815 view simply as Poisson's view. In 1835—surely influenced by Fourier's book—Poisson gives more attention to the perspective Fourier is stressing and mentions various of the things I ascribe to Fourier in this subection, for example the dependence of the decomposition on the domain's shape, see Poisson 1835, §86.

$$v(x,t) = \int_{-\infty}^{+\infty} \frac{F(a)}{2\sqrt{\pi t}} e^{-\left(\frac{a-x}{2\sqrt{t}}\right)^2} \, \mathrm{d}a, \tag{6}$$

as the basic starting point from which all other forms are deduced. Poisson is solving the following problem: within a particular domain, derive the canonical solution formula and then extend that solution to other domains; Fourier, on the other hand, is solving the following problem: across different domains, study how the same physical phenomenon (heat flow) depends on the initial state and shape of the domain. Fourier answers Poisson's objection in the same article:

> The different integrals might be derived from each other, since they are co-extensive. But these transformations require long calculations, and almost always suppose that the form of the result is known in advance.
>
> ...
>
> It has been alleged that in order to solve with certainty problems of this kind, it is necessary to resort in all cases to a certain form of the integral which was denoted as general; and equation Eq. (6) was propounded under this designation; but this distinction has no foundation, and the use of a single integral would only have the effect, in most cases, of complicating the investigation unnecessarily. §428

To appreciate why it would be wrongheaded to treat Eq. (6) as the basic starting point for the other solutions, it will be instructive to compare what, physically, Eq. (6) is saying. Both Fourier's series solution Eq. (3) and Eq. (6) rely on superposition, but what is superposed is different in each case. Fourier superposes 'simple modes', which requires a series of simple modes to describe an initial state. Poisson superposes twice: Eq. (6) itself is a superposition of point sources, it describes the temperature v at a point in space and time as the sum of effects of point-sources defining the initial state $F(a)$. Secondly, to solve the equation on a finite domain, Poisson has to use the method of images; this superposes various initial states to artificially create the effect of a boundary condition (without affecting the initial state *within* the relevant domain). Even apart from the fact that such a deduction is subtle and rather restrictive in its

applicability—the method of images works when the domain tiles the whole space under reflections, as an interval, a box, or a wedge of angle π/n do, but a domain with curved or generic boundary does not—we are forced to resort to things which are physically unreal to deduce something physically real. The issue is not that the simple modes are concrete and the images abstract, both are mathematical abstractions. It is that each mode describes a state the body can actually be put into, one that thereafter evolves 'as if alone existed' and is thereby distinguished §203
'from all other states', whereas the image sources are precisely defined §164
by reference to the space the body does *not* occupy.

Poisson's form puts the initial state center stage and considers the temperature as a sum of the effect of initial point sources; Fourier's form puts the shape of the domain center stage (since it determines ϕ) and considers the temperature as the sum of the effect of simple modes that all evolve independently at their characteristic rate. If, as Fourier would have it, mathematical equations describe phenomena then the mathematical equations have to account for the way shape and initial state bear differently on the movement of heat. The shape of the domain determines which simple modes exist and the rate at which each decays; the initial state determines only how much of each mode is present at the outset. Since all simple modes decay independently at exponential rates, after a relatively short amount of time, only its lowest order term remains; the distribution has become a fingerprint of the geometrical shape of the object, not of its initial distribution of heat. Fourier puts this point thus:

> The forms of bodies are infinitely varied; the distribution of the heat which penetrates them seems to be arbitrary and confused; but all the inequalities are rapidly cancelled and disappear as time passes on. The progress of the phenomenon becomes more regular and simpler, remains finally subject to a definite law which is the same in all cases, and which bears no sensible impress of the initial arrangement.
>
> All observation confirms these consequences. The analysis from which they are derived separates and expresses clearly, (1) the general conditions, that is to say those which spring from the natural properties of heat, (2) the effect, accidental but continued, of the form or state of the surfaces; (3) the effect, not permanent, of the primitive distribution. Preliminary Discourse

Poisson's method has difficulties dealing with the effects of the geometrical domain. The general solution Eq. (6) solves the case that no geometric boundary effect exists. Applying the boundary conditions of a finite bar is possible, but unlike Fourier's simple modes, it requires the use of mathematical artefacts without physical reality. In the case of the sphere, Poisson gets lucky since the ansatz $v = ux$ happens to reduce its PDE to the same form as that of the one-dimensional line. However, Poisson is unable to tackle the case of the cylinder. In the 1835 book Poisson claims to have solved it and refers the readers to his 1815 paper, but no solution to the genuine cylinder problem is found in the paper. Fourier, by contrast, highlights precisely the case of the cylinder, writing:

Poisson 1835, §164 refers back to Poisson 1815 for the cylinder. But what Poisson solves and labels the cylinder there is the one-dimensional ring; he does not address the genuine radial problem involving Bessel functions.

> The most remarkable of the problems which we have hitherto propounded, and the most suitable for showing the whole of our analysis, is that of the movement of heat in a cylindrical body.

§428

Fourier insists on the physical significance of his solution. However, he does *not* consider this a mathematical proof of anything: nowhere in the book does he claim that the physical nature of the phenomenon proves the roots of the transcendental equation are real, even though imaginary roots would correspond to periodically fluctuating temperatures over time rather than the dissipation of heat; nor does he claim that an arbitrary function can be decomposed into such simple modes *because* they have physical significance. Both claims are given mathematical proofs.

See §428 n.9 for Fourier's recognition that imaginary roots cause periodically fluctuating temperatures.

We conclude our comparison of Fourier and Poisson here. Leaving aside his erroneous proof for the general decomposition of a function into a trigonometric series, Poisson is not technically wrong about the mathematics. Fourier himself considered Poisson's deduction of the trigonometric series solution to be (formally) valid. It must therefore be something else that distinguishes Fourier's deduction from Poisson's. Fourier did not name in the abstract what distinguished the two. I want to propose, in terms he does not use and need not have endorsed, what the difference is: Fourier's deduction is a *demonstration* in the Aristotelian sense, Poisson's is not. A deduction is valid when the conclusion follows from the premises; a demonstration must in addition proceed *through the cause*, its middle term exhibiting *why* the conclusion holds. Both Fourier and Poisson reach the same series; but Poisson's middle term—the fundamental solution periodised by fictitious image sources—does not exhibit *why* the solution has the structure it has, and the modes enter his account only secondarily by transforming the integral. Fourier's middle term is the simple modes. In his deduction, the simple modes are derived directly from the PDE by separation of variables; the PDE is in turn derived by accounting for the heat fluxes through the infinitesimal volumes that make up the particular domain and the admissible frequencies μ_n are fixed by the condition at its boundary. The derivation thus exhibits a causal chain from the geometric domain to the simple modes the and the rates at which they decay. In this deduction, one independent premise is required: that arbitrary functions can be decomposed into such simple modes. But in Fourier's work, this premise is an independent result of mathematical analysis, prior to any question about heat, and to be demonstrated through a purely mathematical argument; it is not a consequence of the solution it helps to establish. That priority is what allows it to serve as a premise in a demonstration: the premises of a demonstration must be prior to and better known than the conclusion they support. Poisson deduces the same decomposition from the solvability of the heat problem itself. If it is established that arbitrary functions can be decomposed into the relevant series, then Fourier's derivation is a genuine *demonstration*. Both are valid deductions, but the former affords *epistēmē*: it is the shape from which the modes are deduced, and the modes explain why the distribution takes the form it does. This is the meaning of the title of this thesis: *geometrical shape as a cause.*

Arnold 1983 contains an account of Poisson's proof.

Aristotle's distinction between deduction and demonstration is discussed in the introduction of this thesis.

See the first sentence of §428.

See subsection III.2.

★ # III The generality of Fourier series

Fourier's most famous contribution, as well as most the most controversial, concerns the decomposition of an arbitrary function into a trigonometric series. This includes various issues about the meaning of 'arbitrary function'. In more generic terms there is agreement about what the issue centers on: the definition of the coefficients and the equality between a function and the series. However, a more precise account requires one to spell out, on the one hand, what an 'arbitrary function' is and why the coefficients are well-defined for them, and derivatively, what it means for two functions to be equal. Fourier's conception of 'function' will turn out to be fundamentally different from that of Dirichlet (whose conception of 'function' sets the stage for the rest of the century). Fourier considers a function to be more like a vector, which is itself a consequence of considering a curve as an infinitangular polygon. We will take up his definition of 'function' in subsection III.4.

Fourier's treatment of series decompositions in the *Théorie* is three-tiered: first, in Chapter 3 Sections II–III he gives concrete examples of functions decomposed into a series followed by some remarks; second, in Chapter 3 Section VI he gives more concrete examples and considers the decomposition of an 'arbitrary function': he defines the coefficients and asserts (but does not prove) the generality of the series decomposition followed by some more remarks; third, in Chapter 9, just before his concluding summary of the book, he gives his general proof of the equality of a function and its series (or integral) and gives a deeper perspective on the definition of the coefficients.

Some have considered the fact that he gives most systematic treatment only at the end as evidence that this was an afterthought for him, however, I think this is unfounded. His philosophy for the composition of the book is explicitly spelled out in the preliminary discourse as follows:

Stedall 2008, p.256 takes this perspective.

> [The principles of the theory of heat] could have been explained more concisely by omitting the simpler problems, and presenting in the first instance the most general results; but we wished to show the actual origin of the theory and its gradual progress. When this knowledge has been acquired and the principles thoroughly fixed, it is preferable to employ at once the most extended analytical methods, as we have done in the later investigations.

Preliminary Discourse

In other words, he is following his own development for *pedagogical* reasons. I will follow his three-tiered discussion, but organise each around a specific theme.

III.1 The paradox of odd- and evenness

In Chapter 3, Fourier discusses solves his model problem: a homogenous infinite solid contained between two planes and heated uniformly at one end. He had found a general solution:

$$v = ae^{-x}\cos(y) + be^{-3x}\cos(3y) + ce^{-5x}\cos(5y) + de^{-7x}\cos(7y) + \ldots$$

which are the simple solutions consistent with the BC for y. The problem would therefore be solved if we had, upon setting $x=0$

$$1 = a\cos(y) + b\cos(3y) + c\cos(5y) + d\cos(7y) + \dots.$$

To this end, Fourier constructs an infinite system of linear equations through continual differentiation and setting $y=0$:

$$\begin{aligned}
1 &= a + b + c + d + e + f + \dots\\
0 &= a + 3^2 b + 5^2 c + 7^2 d + 9^2 e + 11^2 f + \dots\\
0 &= a + 3^4 b + 5^4 c + 7^4 d + 9^4 e + \dots\\
0 &= a + 3^6 b + 5^6 c + 7^6 d + \dots\\
0 &= a + 3^8 b + 5^8 c + \dots\\
&\vdots
\end{aligned}$$

He first solves the system formed by the first n equations and then defines how each coefficient changes under an increase of n. He subsequently takes the limit and uses Wallis' product for π to find the values of $a, b, c, \dots$. $\frac{\pi}{2} = \prod_{n=1}^{\infty} \frac{4n^2}{4n^2-1}$
They are given in the solution

$$\frac{\pi}{4} = \cos y - \frac{1}{3}\cos 3y + \frac{1}{5}\cos 5y - \frac{1}{7}\cos 7y + \frac{1}{9}\cos 9y + \dots. \qquad (7)$$ §177

He suggests that the reader can supply a proof that the series converges to $\pi/4$ for each y between 0 and $\pi/2$. He notes how the series equals $-\pi/4$ between $\pi/2$ and $3\pi/2$ and generally that it changes sign when y equals $\pm(2n+1)\pi/2$.

He subsequently derives

$$\frac{x}{2} = \sin(x) - \frac{1}{2}\sin(2x) + \frac{1}{3}\sin(3x) - \frac{1}{4}\sin(4x) + \frac{1}{5}\sin(5x)\dots \qquad (8)$$

when x is between 0 and π and also 'the series given by Euler' $\log(2\cos(x/2)) = \cos(x) - \frac{1}{2}\cos(2x) + \frac{1}{3}\cos(3x) + \dots$. Lastly, he gives a series 'which has not been noticed' §184

$$\frac{\pi}{4} = \sin(x) + \frac{1}{3}\sin(3x) + \frac{1}{5}\sin(5x) + \frac{1}{7}\sin(7x) + \frac{1}{9}\sin(9x)\dots. \qquad (9)$$

for x between 0 and π.

Let us interject that $\pi/4$ has now been decomposed in two distinct ways in Eqs. (7) and (9). But their decomposition holds *only between given bounds.* If one ignored these bounds one would reach the absurd conclusion that $1 = -1$. This introduces Fourier's first substantial contribution: an account of the role of the domain in defining functions and their decomposition into series. A significant part of Fourier's treatment will be centered on clarifying issues about the domain of functions.

In the first tier he notes that Eq. (8) had been known for a long time, 'but the analysis which served to discover it did not indicate why the result ceases to hold when the variable exceeds π' hence 'the origin of the limitation to which each of the trigonometrical series is subject must be sought'. He finds it in an explicit computation of the remainder term. §184
For the concrete case of Eq. (7) he computes a bound for the remainder $\frac{2}{m}(\frac{1}{\cos(x)} - 1)$. Hence as m goes to infinity, the remainder goes to zero for any chosen x between 0 and $\pi/2$. But 'the exactness of this conclusion is based on the fact that the term $1/\cos(x)$ acquires no value which exceeds

all possible limits, whence it follows that the same reasoning cannot apply to the case in which the arc x is not less than $\pi/2$.' He notes that the same analysis can be applied to the other functions.

§188

In the second tier he gives a deeper account of the issue by relating it to the odd- and evenness and the periodicity of functions. I will first explain the issue in my own words and subsequently relate that to what Fourier himself said. Oddness and evenness are of course mutually exclusive, only $f(x)=0$ can be both odd and even. Hence, a paradox emerges: the constant function $f(x)=\pi/4$ is an even function, yet, Eq. (9) equated this supposedly *even* function with a sum of *odd* functions (whose sum is odd), does this imply that a nonzero odd function equals an even function? This paradox emerges if one ignores the domain in the definition of a function. If the constant function is only defined for (some) positive x then it cannot have any property of oddness or evenness. When one extends the function to negative x, one makes a choice whether one extends it to an odd or even function and these give rise to different series.

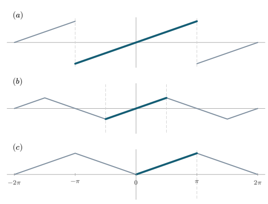
An odd function is a function for which $f(-x)=-f(x)$ and an even function is a function for which $f(-x)=f(x)$. $\sin(x)$ and x^{2n+1} are odd and $\cos(x)$ and x^{2n} are even when $n\in\mathbb{Z}$.

Fourier himself explains the matter by showing that functions containing only even powers can be developped into a series of (odd) $\sin(nx)$. He chooses the case of $\cos(x)$ 'to instance by an example which leaves no doubt as to the possibility of this development' and obtains

$$\cos(x)=\frac{2}{\pi}\left(\left(\frac{1}{1}+\frac{1}{3}\right)\sin(2x)+\left(\frac{1}{3}+\frac{1}{5}\right)\sin(4x)+\left(\frac{1}{5}+\frac{1}{7}\right)\sin(6x)+\ldots\right). \tag{10}$$

for x between 0 and π. He also obtains the three distinct decompositions

$$\begin{aligned}(a):&\quad \frac{1}{2}x=\sin x-\frac{1}{2}\sin 2x+\frac{1}{3}\sin 3x-\frac{1}{4}\sin 4x+\frac{1}{5}\sin 5x-\ldots,\\(b):&\quad \frac{1}{2}x=\frac{2}{\pi}\sin x-\frac{2}{3^2\pi}\sin 3x+\frac{2}{5^2\pi}\sin 5x-\ldots,\\(c):&\quad \frac{1}{2}x=\frac{1}{4}\pi-\frac{2}{\pi}\cos x-\frac{2}{3^2\pi}\cos 3x-\frac{2}{5^2\pi}\cos 5x-\ldots.\end{aligned}$$



Fourier himself does not describe this optionality in terms of a choice of how to extend a function beyond its defined domain. Nevertheless, it is clear that he is aware of the difference between, what Ivor Grattan-Guinness has called, the *algebraic periodicity* and the *geometric periodicity* of a function. Thus, as purely algebraic formulae $\sin(x)$ and x^3 are odd; but as a curve defined between 0 and π they can be periodically extended into an even function, as Fourier's Eq. (10) shows. The understanding of this distinction seems to have originated with Fourier and the lack of understanding by his predecessors and contemporaries seems to have been a major roadblock in the acceptance of Fourier's claims about the generality of trigonometric series. The IBVP dictates whether we seek a decomposition into sines, cosines (or both), yet the IC is totally arbitrary. If the IC were odd but the sought decomposition was even, then that seemed to create a problem. Fourier showed that no such problem exists, though his explanations seem not to have convinced all his detractors. For example, there is a letter from ca.1808–9, probably to Laplace, where he describes the decomposition of a sine into cosines. It includes the

Grattan-Guinness 1990, p.599

remark that 'these theorems are not contrary to the principles of the calculus' which presumably responds to someone who claimed that they were.

Fourier 1808 in Fourier 1980 p.24.

Until now we have considered functions defined for x between 0 and π. But to functions defined for x between $-\pi$ or π the properties of odd and even *do* apply. Such functions are odd, even, or neither. If, for example, the function is odd then it cannot be decomposed into cosines. Hence, in general, a decomposition of $f(x)$ for x between $-\pi$ and π needs a series composed of both sines and cosines. The case that the function is neither odd nor even then represents no problem since, as Fourier recognizes in §233, any function $f(x)$ can be decomposed into an odd and even part as follows

The case of $\pi/4$ for x between 0 and $\pi/2$ leading to Eq. (7) is best understood as being, a priori, equal to $-\pi/4$ between $\pi/2$ and π since the general decomposition into cosines includes the constant $\cos(0x) = 1$ making the decomposition trivial.

$$\phi(x) = \frac{f(x) + f(-x)}{2} \qquad \psi(x) = \frac{f(x) - f(-x)}{2}.$$

$\phi(x)$ is even and $\psi(x)$ is odd and $f(x) = \phi(x) + \psi(x)$.

One last issue remains: $\cos(nx)$ and $\sin(nx)$ are periodic with base period 2π. This period need not align with the domain on which our function is defined. However, this presents no problem since we can simply scale our x appropriately. For example, Eqs. (7) and (9) both have a period of π, but the substitution $x \mapsto \frac{\pi x}{L}$ gives each a period of L instead. Fourier does not hesitate to take the limit of L to ∞, which turns the series decomposition into an integral decomposition (the Fourier transform).

III.2 Arbitrary functions

Considering odd functions only simplifies the length of the calculations but does not affect the nature of the argument; for arbitrary power series one simply gets two systems of linear equations corresponding to the sines and cosines respectively.

The main subject of the second tier is the decomposition of arbitrary functions. First, Fourier generalizes his construction of an infinite system of linear equations to a function given by an arbitrary odd power series. After some clever substitutions he derives

§217

$$\begin{aligned}\frac{1}{2}\pi\phi(x) = \quad & \sin x \ \left\{\phi(\pi) - \frac{1}{1^2}\phi''(\pi) + \frac{1}{1^4}\phi^{(4)}(\pi) - \frac{1}{1^6}\phi^{(6)}(\pi) + \ldots\right\} \\ & -\frac{1}{2}\sin 2x \left\{\phi(\pi) - \frac{1}{2^2}\phi''(\pi) + \frac{1}{2^4}\phi^{(4)}(\pi) - \frac{1}{2^6}\phi^{(6)}(\pi) + \ldots\right\} \\ & +\frac{1}{3}\sin 3x \left\{\phi(\pi) - \frac{1}{3^2}\phi''(\pi) + \frac{1}{3^4}\phi^{(4)}(\pi) - \frac{1}{3^6}\phi^{(6)}(\pi) + \ldots\right\} \\ & -\frac{1}{4}\sin 4x \left\{\phi(\pi) - \frac{1}{4^2}\phi''(\pi) + \frac{1}{4^4}\phi^{(4)}(\pi) - \frac{1}{4^6}\phi^{(6)}(\pi) + \ldots\right\} \\ & + \ldots .\end{aligned} \tag{11}$$

He uses Eq. (11) get a series decomposition for odd polynomials and $e^x + e^{-x}$, which involves the questionable identity $1 - 1 + 1 - 1 + \cdots = \frac{1}{2}$. §218

Now, seeking to 'extend the same results to any function, even to those which are discontinuous and entirely arbitrary', he deduces from Eq. (11) another form of the coefficients. We also have §219

$$a_n = \int_0^\pi \phi(x)\sin(nx)\,\mathrm{d}x \tag{12}$$

for the series

$$\frac{1}{2}\pi\phi(x) = a_1 \sin x + a_2 \sin 2x + a_3 \sin 3x + a_4 \sin 4x + \dots. \tag{13}$$

But this formula for a_n is *also* the formula for the area contained by the product curve $\phi(x)\sin(nx)$ and the three axes $y = 0$, $x = 0$ and $x = \pi$. It is therefore defined for 'all possible functions' since the integral corresponds to a specific area 'whether their character can be expressed by known methods of analysis, or whether they correspond to curves traced arbitrarily'. In §221 he 'verifies' this result by obtaining the definition of §224
a_n directly, via the orthogonality of the sines with different frequencies. Given Eq. (13), he works out the concrete computations wich show that multiplying of both sides by $\sin(nx)\,\mathrm{d}x$ and integrating between 0 and π gives Eq. (12).

After applying his new coefficient formula to functions whose series we had already obtained by other means, he turns to a set of examples of 'discontinuous' functions expressed by a series. These address another objection to his series. His first example is

> the case in which the function to be developed has definite values when the variable is included between certain limits and equals nul when the variable is included between other limits. We stop to examine this particular case, since it is presented in physical questions which depend on partial differential equations, and was proposed formerly as an example of functions which cannot be developed in sines or cosines of multiple arcs. Suppose then that we have reduced to a series of this form a function whose value is constant, when x is included between 0 and α, and all of whose values are nul when x is included between a and π. §226

Observe how he supposes the decomposition has taken place; he subsequently applies the coefficient formula. We get

$$\frac{\pi}{2}\phi(x) = \Big\{(1-\cos(\alpha))\sin(x) + \frac{1-\cos(2\alpha)}{2}\sin(2x) \\ + \frac{1-\cos(3\alpha)}{3}\sin(3x) + \dots\Big\} \tag{14}$$

Throughout §§226–8 he similarly considers

(i) $f(x) = \sin(\pi x/\alpha)$ on $(0,\alpha)$ and 0 on (α,π);
(ii) $f(x) = (\pi/2)^2 - x^2$ on $(0,\pi/2)$ and 0 on $(\pi/2,\pi)$;
(iii) $f(x) = x$ on $(0,\alpha)$, α on $(\alpha,\pi-\alpha)$, and $\pi - x$ on $(\pi-\alpha,\pi)$.

These are 'discontinuous' functions in Euler's sense: their behavior on one part of the domain entails nothing about their behavior elsewhere; they are not subject to a common law. The modern relative of Eulerian continuity is roughly the property of analyticity. Fourier is showing us that a function representable by a trigonometric series needn't be continuous (or analytic). The point of naming analyticity is that we can appreciate, in retrospect, how Fourier's work on 'discontinuous' functions is an essential step in the growing separation of real and complex analysis: in the former functions with trigonometric expansions are central, whereas analytic functions are central to the latter.

Having established that there exist discontinuous functions that can be expressed by a series, we turn to formulating Fourier's general theorem:

that we can 'develop any function whatever in a series formed of sines and cosines of multiple arcs', resulting in the equality §226

$$F(x) = \sum_{n=0}^{\infty} a_n \cos(nx) + b_n \sin(nx) \tag{15}$$

with

$$a_n = \frac{1}{\pi} \int_{-\pi}^{+\pi} F(x) \cos(nx)\, \mathrm{d}x \qquad b_n = \frac{1}{\pi} \int_{-\pi}^{+\pi} F(x) \sin(nx)\, \mathrm{d}x. \tag{16}$$

He is very aware that this requires the convergence of Eq. (15). Writing,

> In general, the trigonometric series at which we have arrived, in developing different functions are always convergent, but it has not appeared to us necessary to demonstrate this here; for the terms which compose these series are only the coefficients of terms of series which give the values of the temperature; and these coefficients are affected by certain exponential quantities which decrease very rapidly, so that the final series are very convergent. With regard to those in which only the sines and cosines of multiple arcs enter, it is equally easy to prove that they are convergent, although they represent the ordinates of discontinuous lines.

i.e. for $t = 0$

The quote continues with Fourier clarifying the meaning of convergence:

> This does not result solely from the fact that the values of the terms diminish continually; for this condition is not sufficient to establish the convergence of a series. It is necessary that the values at which we arrive on increasing continually the number of terms, should approach more and more a fixed limit, and should differ from it only by a quantity which becomes less than any given magnitude: this limit is the value of the series. We may prove rigorously that the series in question satisfy the last condition. §228

This should satify the reader till the end of the book, where he at last gives a proof. Note that he clearly does not consider the derivation of the coefficient formula Eq. (16) to constitute a proof: his usage of 'verifier' as against 'proof' has to be deliberate and in the case of the discontinuous function he felt the need to assert the possibility of a decomposition before using the formula. If Fourier had stopped here, then he would wholeheartedly deserve the reputation often ascribed to him—someone with great intuition but who lacked rigor. But, echoing the view of British mathematician George Gibson in a 1893 paper *'On the History of the Fourier Series'*:

> in this respect Fourier has received less than justice. No doubt, the investigations of Chapter III can hardly be accepted as doing more than suggesting the truth of the general theorem, but it is different with those of Chapter IX.

Gibson 1892, p. 147. He explicitly named Riemann's account as the object of criticism.

Let us therefore turn to those investigations, which I have referred to as the third tier.

III.3 General perspective on the meaning of Fourier's series

The third tier of Fourier's treatment contains his general proof of the development of arbitrary functions and, alongside it, a clarification of the coefficient formula. Both will involve the more genreral case of

$$a_1 \phi(\mu_1 x) + a_2 \phi(\mu_2 x) + a_3 \phi(\mu_3 x) + \ldots \tag{17}$$

where 'the function ϕ [...] may be very different from a sine. The numbers $\mu_1, \mu_2, \mu_3 \ldots$ instead of being integers, are given by a transcendental equation, all of whose roots infinite in number are real.' He takes up the proof of generality first, but we will start with the coefficient formula, because it will draw out what he means with 'function' in practice. We then discuss his conception of function in subsection III.4 and conclude with his proof in subsection III.5.

Given ϕ_i and μ_i as in Eq. (17), given the coefficient formula

$$a_i = \rho_i \int_0^X \phi(\mu_i x) f(\alpha) \sigma_i \, \mathrm{d}\alpha$$

where σ_i and ρ_i are appropriate weighing factors, and given that any function can be developed as Eq. (17),

> the process which determines the coefficient a_i does not explain why all the roots must enter into equation Eq. (17), nor why this equation refers solely to values of x, included between 0 and X.
>
> To answer these questions clearly, it is sufficient to revert to the principles which serve as the foundation of our analysis. §424

He divides X into an infinite number of n parts and for the f_i corresponding to $i\mathrm{d}x$, 'this arranged, the equation Eq. (17) represents n equations of the first degree'. Granting modern matrix notation, he sets up the system of equations

$$\begin{pmatrix} \phi(\mu_1 \ \mathrm{d}x) & \phi(\mu_2 \ \mathrm{d}x) & \cdots & \phi(\mu_n \ \mathrm{d}x) \\ \phi(\mu_1 2\,\mathrm{d}x) & \phi(\mu_2 2\,\mathrm{d}x) & \cdots & \phi(\mu_n 2\,\mathrm{d}x) \\ \phi(\mu_1 3\,\mathrm{d}x) & \phi(\mu_2 3\,\mathrm{d}x) & \cdots & \phi(\mu_n 3\,\mathrm{d}x) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(\mu_1 n\,\mathrm{d}x) & \phi(\mu_2 n\,\mathrm{d}x) & \cdots & \phi(\mu_n n\,\mathrm{d}x) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_n \end{pmatrix}. \tag{18}$$

Subsequently, he spells out how to determine a_i

> To determine the first coefficient a_1, we multiply the first equation by σ_1, the second by σ_2, the third by σ_3, and so on, and add together the equations thus multiplied. The factors $\sigma_1, \sigma_2, \sigma_3, \ldots \sigma_n$ must be determined by the condition, that the sum of all the terms of the second members which contain a_2 must be nothing, and that the same shall be the case with the following coefficients $a_3, a_4, \ldots a_n$. All the equations being then added, the coefficient a_1 enters only into the result, and we have an equation for determining this coefficient. ... Now it is evident that this process of elimination is exactly that which results from integration between the limits 0 and X.
>
> We see by this that the process which serves to determine these coefficients, differs in no respect from the ordinary process of elimination in equations of the first degree. The number n of equations is equal to that of the unknown quantities $a_1, a_2, a_3 \ldots a_n$, and is the same as the number of given quantities $f_1, f_2, f_3 \ldots f_n$. The values found for the coefficients are those which must exist in order that the n equations may hold good together, that is to say in order that equation Eq. (17) may be true when we give to x one of these n values included between 0 and X; and since the number n is infinite, it follows that the first member $f(x)$ necessarily coincides with the second, when the value of x substituted in each is included between 0 and X. §424

A system of n equations in n unknowns; multiplication factors chosen to make all coefficients but one disappear; solvability referred to the match between the number of equations and the number of unknowns, these are Fourier doing linear algebra before a systematic theory of linear algebra. What stands out is the identification: 'this process of elimination is exactly that which results from integration between the limits 0 and X.' For Fourier, the integral formula for the coefficients *is* the solution of the system, not a separate analytic formula. One issue must be flagged: what does it mean that n is infinite? We can follow the reasoning for finite n and defer it to subsection III.4.

The abstract exposition of this process explains two things. It shows that

In my exposition I swapped the order in which Fourier discusses these two things.

> [it] is certain that on giving to x any value whatever included between 0 and X, the second member of this equation becomes equal to $[f(x)]$; this is a necessary consequence of our process. But it nowise follows that on giving to x a value not included between 0 and X, the same equality would exist. §426

This is his deepest explanation for why equality between a function and its series only holds on the domain of the function.

Moreover, it explains why we cannot omit 'one or more terms which corresponds to one or more roots μ_i'. If we did so then

(ϵ) equates a function with its Fourier series (using the coefficient formula); (e) merely describes the Fourier series.

> equation (ϵ) would not in general be true. To prove this, let us suppose a term containing μ_j and a_j not to be written in the second member of equation (e), we might multiply the n equations respectively by the factors
>
> $$\mathrm{d}x \sin\left(\mu_j \,\mathrm{d}x\right), \mathrm{d}x \sin\left(\mu_j 2 \,\mathrm{d}x\right), \mathrm{d}x \sin\left(\mu_j 3 \,\mathrm{d}x\right) \dots \mathrm{d}x \sin\left(\mu_j n \,\mathrm{d}x\right);$$
>
> and adding them, the sum of all the terms of the second members would be nothing, so that not one of the unknown coefficients would remain. The result, formed of the sum of the first members, that is to say the sum of the values $f_1, f_2, f_3 \dots f_n$, multiplied respectively by the factors
>
> $$\mathrm{d}x \sin\left(\mu_j \mathrm{d}x\right), \mathrm{d}x \sin\left(\mu_j 2 \mathrm{d}x\right), \mathrm{d}x \sin\left(\mu_j 3 \mathrm{d}x\right) \dots \mathrm{d}x \sin\left(\mu_j n \mathrm{d}x\right),$$
>
> would be reduced to zero. This relation would then necessarily exist between the given quantities $f_1, f_2, f_3 \dots f_n$; and they could not be considered entirely arbitrary, contrary to hypothesis. §425

Consequently, 'if $f(x)$ had all the generality possible,' then

> we ought to suppose the second member of equation (e) to be complete, and the investigation shows what terms ought to be omitted, since their coefficients become nothing. §425

He is giving a mathematical argument as to why no term of the series can be dropped a priori: if a term was dropped, then multiplying by the corresponding missing function and integrating gives 0 on the RHS so that the data $f_1, f_2, \dots f_n$ are forced to satisfy a linear relation. Using later vocabulary, if the system were incomplete, some nontrivial linear functional would annihilate it, and arbitrary data could not be represented. Grattan-Guinness credits Fourier with having 'stumbled across the completeness problem (as we now say)', but holds that he 'never had an idea of how to solve the problem detected' and 'only gave it one small airing in print'. However, is Fourier considering himself as spotting a *problem* or is this supposed to be an *explanation*: why must all the roots

Grattan-Guinness 1990, p. 598

enter? because otherwise the data are constrained. It becomes an open problem only when 'function' is interpreted as 'real function' rather than as the sequence $f_1, f_2, f_3 \dots f_n$, for then the n values no longer exhaust the data.

Fourier has thus clarified two issues—why every root must enter, and why the equality holds only between 0 and X—by reverting to 'the principles which serve as the foundation of our analysis': a function is defined by n given quantities $f_1, f_2, \dots f_n$, exactly as many as the system has unknowns. However, he considers n to be infinite, what does this mean? And, more broadly, what does the practice exhibited in this proof imply for the definition of 'function'?

III.4 Fourier's definition of 'function'

Fourier answers both questions in the closing paragraphs of the *Théorie*, where the 'arbitrary functions' referred to throughout the book are at last taken up systematically and where he gives various characterizations of them. One that is commonly cited is the following: §415–427

> In general the function $f(x)$ represents a succession of values or ordinates each of which is arbitrary. An infinity of values being given to the abscissa x, there are an equal number of ordinates $f(x)$. All have actual numerical values, either positive or negative or nul. We do not suppose these ordinates to be subject to a common law; they succeed each other in any manner whatever, and each of them is given as if it were a single quantity. §417

Fourier thus defines a function as an *infinite succession* of values (corresponding to another succession of values). It is 'arbitrary' when the values are subject to no common law: knowing some of them entails nothing about the others. A 'continuous' function, by contrast, is a succession whose values do stand in a common relationship, one expressed by an analytic formula.

Here 'continuous' means the Eulerian sense.

Addressing the audience of his time, Fourier's main emphasis lies on distinguishing the properties of arbitrary functions from those of continuous functions. A modern reader, however, can easily fall into the trap of thinking that Fourier essentially means a real function, i.e. an arbitrary map from the real numbers to the real numbers (with additional properties taken for granted in his actual practice). However, when Fourier talks about a 'succession of values', I take him to mean a *sequence of* of values. This is how he reasons with 'function' throughout the *Théorie*, in particular in the argument discussed in subsection III.3. If a function maps a *sequence* to another *sequence*, then it does not assign values to all points of a continuous domain. The points of a continuous domain cannot be put in a sequence, they are uncountable. The concept of a real function is therefore fundamentally different from what Fourier has in mind. The real function was introduced by Fouriers pupil Dirichlet who described an arbitrary function thus:

See also §202 and the start of Chapter 9.

> Let $\phi(x)$ be a function of x, having a finite and determined value for every value of x comprised between $-\pi$ and π. Dirichlet 1829

Despite the seeming similarity to Fourier's words, Dirichlet means something fundamentally different. With 'every value of x' Dirichlet had in mind we today call 'any real number'. This is clear from the pathological example he gives of a function which cannot be represented by a Fourier series:

> $\phi(x)$ [is] equal to a determined constant c when the variable x obtains a rational value, and equal to another constant d when this variable is irrational.

This function cannot be defined as a sequence corresponding to another sequence and therefore falls strictly outside Fourier's definition of 'function'. Fourier's 'function' has a *countable* domain whereas Dirichlet's 'function' has an *uncountable* domain. Conversely, Fourier explicitly considers the so-called Dirac delta distribution to be a function. This cannot be defined as a real function. The two concepts are therefore neither related as the informal to the precise, nor as the narrower to the wider: neither extends the other.

Jesper Lützen, in Jahnke 2003, p. 157, seems to consider them to be equivalent, writing that Dirichlet 'accepted' Fourier's definition.

A glaring issue with Fourier's words is that it seems to commit him to the existence of infinite and infinitesimal quantities. Arguments involving infinitesimals had been used rather informally from the early calculus onwards, were still common in Fourier's time but were abandoned in the 'age of rigor' in favor of arguments involving limits. Does the rejection of infinitesimals require us to leave Fourier's definition behind? Does it explain why the concept of a real function became the basic concept of analysis instead? Fourier does not consider his usage of the infinite to imply the existence of actually infinite (or infinitesimal) quantities, calling n 'infinite' is just a shorthand for the continual increase of n. His proofs involving the infinite always consist of setting up a system of n quantities first and then declaring n to be infinite, they

> [suppose] that notion of infinite quantities which has always been admitted by geometers. It would be easy to offer the same proof under another form, examining the changes which result from the continual increase of the factor n. §417

Fourier thus has an Aristotelian conception of infinity as the *potentially* infinite: 'these n quantities have values *actual*, and consequently not *infinite*.' If a quantity were infinite then it could not be actual. So much for Fourier's abstract philosophical commitments. But are they coherent with his actual practice? Can Fourier's definition be made intelligible without commiting oneself to the existence of actual infinities?

§426

Aristotle 1984: "Infinity turns out to be the opposite of what people say it is. It is not 'that which has nothing beyond itself' that is infinite, but 'that which always has something beyond itself." Phys. III.3 207a.

We will now elaborate on Fourier's definition in the form of four comments: (1) its historical origins and similarity to Laplace's view of 'arbitrary function'; (2) the scope of Fourier's definition; (3) whether it can be formalized without invoking the actually infinite and its relationship to the 19th century rigorization of analysis; and (4) the fundamental difference between Fourier and his successors which underlies their different conceptions of 'function'.

(1) Fourier's definition has its roots in the basic ideas of the Leibnizian calculus. Leibniz' work in the calculus originated from the study of discrete number sequences. Considering the sequence $y_1, y_2, y_3, \ldots$, he had defined two operations for such sequences: '$\int$ signifies sum, d difference'.

Applied to a sequence y_n, these define the sequence of differences $(y_2 - y_1, y_3 - y_2, y_4 - y_3 \ldots)$ and the sequence of sums $(y_1, y_1 + y_2, y_1 + y_2 + y_3, \ldots)$ respectively. The calculus emerged from applying these ideas to geometry, to the study of continuous magnitudes. Let us recall one central problem which motivated the calculus: quadrature, or the determination of areas. The problem of course concerns the area of figures with continuously curved boundaries, like a circle; the area of rectilinear figures presented no difficulty in principle, being reducible to triangulation. The basic idea for general quadrature, from the Greeks onwards, had been to take a sequence of polygons so that the polygons coincided more and more with the curve containing the area as the number of sides increased. If one could determine the area of the sequence of polygons, then one would know the area of the figure. The crucial Leibnizian insight consists of considering a curve as an infinitangular polygon and extending his calculus of finite difference sequences to infinitesimal difference sequences. Leibniz's ideas were systematized by his disciple Johann Bernoulli whose work was made into a textbook by the Marquis de l'Hôpital. Consider one of its postulates:

Leibniz 1675, pp.292–293

> Postulate II: We suppose that a curved line may be considered as an assemblage of infinitely many straight lines, each one being infinitely small, or (what amounts to the same thing) as a polygon with an infinite number of sides, each being infinitely small, which determine the curvature of the line by the angles formed amongst themselves.

de l'Hôpital 1696, §3

Leibniz's genius consisted of working out the basic rules and procedures of the number sequences in the case that $\mathrm{d}x$ is infinitely small. The calculus reduces quadrature to the determination of $\sum y_n \, \mathrm{d}x_n$ when $\mathrm{d}x_n$ is infinitesimal, for which we use the special notation $\int y \mathrm{d}x$. Observe that expression literally is a sum of rectangular areas $y \times \mathrm{d}x$. The bounds of integration are dictated by the nature of the area. Similarly, the determination of tangets is reduced to the determination of $\mathrm{d}y / \mathrm{d}x$ and the rectification of lengths is reduced to the determination of $\int \sqrt{\mathrm{d}y^2 + \mathrm{d}x^2}$.

The concept of 'function' did not play a big role in Leibniz' own work. His calculus centered on the idea of a variable, i.e. a changing magnitude. If x and y are both variables (e.g. describing the abscissa and ordinate of a circle) then $\mathrm{d}y$ and $\mathrm{d}x$ go hand-in-hand and imply each other, no distinction between 'dependent' and 'independent' variable needs to be made. The further development of higher order differentials (whose value depends on whether x or y is the independent variable) and multivariable functions ($\mathrm{d}z$ is ambiguous if z is a function of x and y) made clear the fundamental need to distinguish *dependent* and *independent* variables, 'function' (as against 'variable') embodies this difference. Euler's textbooks gave 'function' the central role it has had ever since.

See Bos 1974 for Leibniz' ideas.

Fourier ideas inherit Leibniz' classical perspective, his 'function' is the analytical counterpart of the infinitangular polygon, which itself—by the above postulate—corresponds to any curve. Fourier differs from Leibniz in that he does not take the infinite to be actual. Similarly, 'infinitangular polygon' is just a shorthand for considering a sequence of polyons with more and more sides. This philosophical difference does not affect the technical content of the Leibnizian calculus (as Leibniz himself knew and insisted on), but it will be crucial when we consider, in comment (3), if Fourier's definition can be formalized.

Additionally, both Fourier and Leibniz knew that analysis applies to magnitudes more generally—to time, mass, and for Fourier, even to temperature and magnitudes found in probability—but a conversion of one kind of a magnitude to another does not change the purely quantitative relationship which is studied in analysis. Leibniz represented masses with line segments and similarly Fourier represents temperature distributions by means of geometric surfaces, but this merely transforms the sought quantitative relationship, it does not identify it. For Fourier, as well as the Leibnizians, it is important that each quantity has a specific dimension and Fourier actually gives the first systematic treatment of dimensional analysis. But, for Fourier, dimensional considerations are merely the anteroom of analysis; determining the relevant analytic equations and physical constants numerically, '[reduces] the physical questions to problems of pure analysis, and this is the proper object of theory' so that now 'all the problems relating to the propagation of heat depend only on numerical analysis.'

e.g. in §3 and §99.

Fourier does dimensional analysis in Chapter 2 Section VIII of the *Théorie*.

Preliminary Discourse

§11

Laplace's perspective on 'arbitrary function' is the same as Fourier's. In Chapter 1, we saw him use 'arbitrary function' to refer to a function defined by an arbitrary power series (i.e. 'continuous' in the Eulerian sense). This, however, seems to be a bad habit which contradicts his explicit views on the matter. In a 1779 paper *Mémoire sur les Suites* Laplace argued that we should understand the infinitesimal calculus as an extension of the calculus of finite differences under the supposition that the finite difference becomes infinitesimal, repeating the idea that originally birthed Leibniz' calculus. This perspective leads him to conclude that 'discontinuous' functions ought to be admitted into analysis. He explains that to understand what an arbitrary function is, 'it is necessary to consider it in the finite differences, and to observe that which arrives in the passage from these differences to the infinitely small differences.' For $y_{x,x'}$, he sets up the following table as the general solution to a finite difference equation

Laplace 1779

Laplace 1812 Bk.I Ch.2 pp.77–79

$$\begin{array}{lllllll}
y_{0,0}, & y_{1,0}, & y_{2,0}, & y_{3,0}, & \cdots & y_{n-1,0}, & y_{n,0}, \\
y_{0,1}, & y_{1,1}, & y_{2,1}, & y_{3,1}, & \cdots & y_{n-1,1}, & y_{n,1}, \\
y_{0,2}, & y_{1,2}, & y_{2,2}, & y_{3,2}, & \cdots & y_{n-1,2}, & y_{n,2}, \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
y_{0,\infty}, & y_{1,\infty}, & y_{2,\infty}, & y_{3,\infty}, & \cdots & y_{n-1,\infty}, & y_{n,\infty},
\end{array}$$

and then considers what happens when n becomes infinite, so that the domain (of finite length) is divided into infinitesimal parts dx. In that case, the first row corresponds to an 'arbitrary function'. However, Laplace does impose some condition for the employment of arbitrary functions in the theory of PDEs. If the function were to have a finite jump for an infinitesimal difference dx then the derivative dy/dx would be infinite. Consequently, Laplace asserts that if the PDE is of order n then the function cannot have a finite jump between two values in any difference smaller than the nth partial difference. Fourier differs from Laplace in never stating such a condition.

Laplace 1812, p.80

(2) What functions fall under Fourier's definition? Of course any 'continuous function' defined by an analytical equation does, since its values can simply be determined by computation from specific values for the independent variable. Moreover, so do various functions which are 'discontinuous', like a square wave. More interestingly, as alluded to, Fourier considers the Dirac delta to be a function:

> The expression $\frac{1}{2}+\sum_{i=1}^{\infty}\cos(i(x-\alpha))$ represents a function of x and α, such that if it be multiplied by any function whatever $F(\alpha)$, and integrated with respect to α between the limits $\alpha=-\pi$ and $\alpha=\pi$, the proposed function $F(\alpha)$ becomes changed into a like function of x multiplied by the semi-circumference π.

§235. Note that Fourier's proof that arbitrary functions have Fourier series essentially comes down to proving this quote, i.e. proving that that expression has the properties he describes, see subsection III.5.

And he also uses it in the form of an infinite spike of infinitesimal width:

> In the preceding problem we have determined the variable state of an infinite prism a definite portion of which was affected throughout with an initial temperature f. We suppose that the initial heat was distributed through a finite space from $x=0$ to $x=b$. We now suppose that the same quantity of heat bf is contained in an infinitely small element, from $x=0$ to $x=\omega$. The temperature of the heated layer will therefore be $\frac{fb}{\omega}$.

§388

In other words, the initial state is taken to be a Dirac delta and is plugged into the solution to the heat equation.

So much for functions which Fourier himself explicitly considered. However, there is also an important class of functions that satisfies Fourier's abstract characterisation of 'function', but which he (most likely) was not aware of. Its representative is the Weierstrass function, a continuous nowhere differentiable function. The evidence that he probably wasn't aware of such functions is his explicit statements in manuscripts that all curves have a finite arclength. The Weierstrass function can be interpreted geometrically as a curve, but its arclength is infinite. The broader class of functions which the Weierstrass function represents is (continuous) functions of unbounded variation. Are these then part of Fourier's conception of 'function' or not? It seems to me that both 'yes' and 'no' are not satisfactory answers. It is possible to grasp a concept and even define it, without grasping the diversity of things falling under it. Whoever knows the rules of chess implicitly possesses the entire class of legal positions; yet that class contains positions no player would imagine and that are not obviously possible. Similarly, in assessing the scope of Fourier's definition, we must distinguish what the definition entails from what he *thought* it entailed. Formal rigor exposes such gaps to the practitioner. Fourier's definition lacks the rigor that the nineteenth century would come to demand, and so the gap could go unnoticed; he did not confront the contradiction between his definition and his belief in all curves having finite arclength. The irony is of course that the function that exposes the limits of Fourier's conception is constructed using his own ideas: the Weierstrass function is itself defined by a trigonometric series!

Fourier n.d.(a). The context makes clear that it means: curves within a finite region.

(3) Fourier considers taking N to be infinite equivalent to letting N increase indefinitely. But what is the meaning of the sequence $f_1, f_2, f_3 \cdots f_N$ under the continual increase of N? The definition of a real function seems to sidestep the issue since it does not make an explicit reference to infinities. However, it has merely transformed the issue since the obvious question now is: what is a real number and what does it mean that the domain consists of 'all' real numbers? Fourier did not address this issue, nor did the formalization of the limit concept settle it.

Geometrically, as N increases, the polygon coincides more and more with the curve so that when N is increases without limit, the curve becomes indistinguishable from the polygon. The properties of the curve are then to be determined through the properties of the polygon. Both the curve and the polygons are geometric objects; 'function', on the other hand, is something algebraic, it pertains to a sequence of *numbers* (that measure the ordinates of the polygon). If we distinguish the geometric polygon from the sequence of numbers, then what does $N \to \infty$ mean for the sequence of numbers? Does it get closer and closer to an infinitely long sequence? If one holds that geometric objects like polygons and circles exist, does that imply that there exist infinitely long sequences of numbers to which the finite sequences tend? The exact same problem exists in a more basic form for magnitudes and numbers, which is why adopting 'real function' instead does not sidestep the issue. Given a circle and a sequence of inscribed polygons, the area the polygons tends towards the area of the circle as the number of sides increases; but does that mean that the rational *number* x_n—describing the area of a polygon—tends towards a *number* x—describing the area of the circle? If the number x is itself defined as the limit of x_n then its definition is circular. This apparent circularity is the problem of defining the real numbers.

The modern formalization of the real numbers does not consider them to be things which exist independent of the rational numbers and towards which sequences of rational numbers tend. Instead, the real numbers are constructed as equivalence classes of *sequences* of rational numbers. Two sequences x_n and y_n of rational numbers are equivalent if for any $\varepsilon > 0$ there is an N so that $|x_n - y_m| < \varepsilon$ for $n, m \geq N$. The equivalent sequences share a property: their difference can be made smaller than any given magnitude; the real number *is* the sequence *qua* that shared property. Real numbers like π are not things towards which sequences of rational numbers tend, rather, real numbers are something which sequences of rational numbers share. π is not what 3.14159... yearns to be, it is what the forms 3.14159..., $[3;7,15,1,292,1,\dots]$, $(11.00100\dots)_2$ have in common *qua indefinitely extendible sequences of rational numbers*. In other words, the measurement π—a measurement of infinite resolution—is *defined* in terms of indefinite sequences of finite measurements, as an equivalence class of such sequences, since there are different ways of measuring the same magnitude.

These forms refer to the decimal expansion, continued fraction and binary expansion respectively.

Put that way, the analogous idea for Fourier's 'function' as an infinite succession of values would be to define a function as a *sequence of finite number-sequences* (corresponding to a sequence of polygons), each finite number-sequence measuring the ordinates of one polygon. To make this work, one needs a condition playing the role that the Cauchy criterion plays for sequences of rationals. No such formalization is found in Fourier's work; its essential inspiration is the construction of the real numbers, which came half a century later. In Chapter 3 we state such a condition and show that it formalizes Fourier's 'infinite succession of values' in analogy with the formalization of the real numbers. Fourier's proof that any 'arbitrary function' can be developed in a trigonometric series then becomes a theorem about this kind of function (while exposing the need for additional clarification of the statement).

Analysis did not develop along these lines; the real function became central instead. This creates an asymmetry: real numbers are constructed from sequences of increasing finite resolution, but a real function has, by definition, infinite resolution on its domain. The asymmetry is what makes room for the Dirichlet function encountered above: a function which cannot be obtained from any sequence of finite number-sequences, and which makes sense only if one takes for granted, from the start, the ability to distinguish every real number from every other.

The point of these considerations is the following. The rigorization of analysis is usually told as a story in which a definition like Fourier's could not survive: the 'infinite succession of values' was a pre-rigorous manner of speaking, to be replaced by the real function once standards tightened. But, since the method which rigorized 'number' applies equally to Fourier's 'function', the demand for rigor does not logically require the real function. Nothing in the requirement of precise definitions and proofs demands a specific definition of 'function'. What does distinguish the two conceptions of 'function' lies in the conception of the domain itself, which we take up in the next comment. The real function has no means of expressing resolution on the domain; two functions differing at a single point are simply distinct. The notion of a resolution re-enters the theory only later in the theory of distributions, where a function is characterized only through its pairings against test functions.

e.g. Vretblad 2003 motivates distributions by analogizing test functions with resolutions of a physical measurement.

(4) I have argued that there is a fundamental difference between Fourier's and Dirichlet's conception of 'function'. However, 'function' is not fundamental, it presupposes a domain; the two definitions correspond to two different conceptions of the domain, two different conceptions of the geometric continuum.

Fourier's conception of the continuum is explicitly formulated by him in manuscripts on the foundations of geometry:

> Geometry presupposes the notion of space and of *its parts*. A volume is a part of space. When a given volume is divided into two parts, that which is common to these two parts is a *surface* (the two parts are said to be *contiguous*). When a surface is divided into two parts, that which is common to the two contiguous parts is a line. When a line is divided into parts, that which is common to two contiguous parts dividing it is a point. [*Marginal addendum*:] A surface terminates a volume, a line terminates a surface. A point is an extremity of a line.

Fourier n.d.(a), fol.21 [5]; compare: fol.4r and fol. 8–10, copied at fol. 17–20; their dating is discussed in Charbonneau, Catalogue.

What is primitive is extension and its part–whole structure. Surfaces, lines and points are *boundaries*, not constituents of space, and a boundary exists only where a division has been made: it is 'that which is common' to two contiguous parts. A point is accordingly not a part of a line—the parts of a line are lines—but the extremity at which two of its parts meet. This is the classical, Aristotelian perspective on the continuum: divisible without end and not composed of points. The points 'contained in' a line exist potentially, a division actualizes finitely many of them and the divisions can be refined without end, but they never form a completed totality. The continuum as a whole is given; a point exists only where a division has been made, as the Fourier passage above says explicitly.

Cf. *Théorie* §428: 'We do not know that anyone has been able to imagine the movement of heat as being produced in the interior of bodies by the simple contact of the surfaces which separate the different parts. For ourselves such a proposition would appear to be void of all intelligible meaning. A surface of contact cannot be the subject of any physical quality; it is neither heated, nor coloured, nor heavy.'

Aristotle 1984: 'nothing that is continuous can be composed of indivisibles: e.g. a line cannot be composed of points, the line being continuous and the point indivisible.' Phys. VI.1, 231a.

Fourier's definition of 'function' has to be understood in light of such a view of the continuum. On this view, one cannot stipulate values 'at every point of the line', because the points of the line are not a given totality. What one *can* do is divide the domain into n contiguous parts and assign to the division the values $f_1, f_2, \dots f_n$ and then refine, declaring n 'infinite' in the sense of its continual increase.

The real function makes the point, instead of the line, basic. Dirichlet's 'for every value of x comprised between $-\pi$ and π' refers to the points of the interval as a completed totality, each member of which possesses its identity individually and prior to any division. Setting $\phi(x) = c$ on the rationals and d on the irrationals means discriminating between points that no division will ever separate since, within every part however small, both kinds of points exist. Such a function can only be defined on a continuum whose points are actual. Consequently, the line *is* a set of points, and continuity—the primitive characteristic of extension itself for Aristotle and Fourier—becomes a derived property that a point-set may possess or lack. The arithmetical counterpart of this continuum is the set $\mathbb{R}$: the points of the line are identified with the real numbers, and the continuity of the line becomes the completeness of the set.

The difference between Fourier's and Dirichlet's 'function' is thus explained by two different conceptions of the continuum. Since the arithmetization of analysis in the nineteenth century, the point-set continuum has tacitly become the only one, but it would be a mistake to read it back into Fourier.

III.5 Fourier's general theorem

At last we turn to Fourier's proof that an arbitrary function admits the kind of decomposition necessary to solve the heat equation. He first establishes it for the integral with the standard trigonometric components $\sin(nx)$, $\cos(nx)$ (§§415–17); afterwards he explains how the argument applies to trigonometric series (§§418–19, §423); and he also asserts that the argument applies to a more general class of decompositions (§423). The underlying argument is structurally identical in each case, so we will consider the first and comment on the general case at the end.

We seek to prove the equality

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} f(y)\cos(\xi y)\,\mathrm{d}y\right]\cos(\xi x)\,\mathrm{d}\xi$$
$$+\frac{1}{2\pi}\int_{-\infty}^{+\infty}\left[\int_{-\infty}^{+\infty} f(y)\sin(\xi y)\,\mathrm{d}y\right]\sin(\xi x)\,\mathrm{d}\xi. \quad (19)$$

§415 will take off from a rewritten form, derived by combining the two integrands and applying the cosine subtraction formula, resulting in

$\cos(a-b) = \cos(a)\cos(b) + \sin(a)\sin(b)$

$$\frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(y)\cos(\xi(y-x))\,\mathrm{d}y\,\mathrm{d}\xi \quad (20)$$

Subsequently, Fourier swaps the integrals and integrates with respect to ξ to obtain

$$\frac{1}{\pi}\int_{-\infty}^{+\infty} f(y)\frac{\sin(\xi(y-x))}{y-x}\,\mathrm{d}y, \quad (\xi = \infty). \quad (21)$$

Swapping integrals without justification invites the scrutiny of modern readers, but in this case it is more of a symbolic than a logical defect. When Fourier spells out what he means with Eq. (21) it becomes clear that the underlying reasoning is fine. Fourier has no 'lim' symbol to indicate that the order of taking the limits for both bounds of integration has not been swapped, forcing him to simply say 'ξ is infinite', which does not distinguish the order. This symbolic defect has logical significance, because it means his notation cannot express such differences and we cannot know to what extent he was aware of them. This, generally, is a defect of infinitesimals. Fourier considers their use to be logically equivalent to limit arguments, but symbolically they are less expressive since 'n and m are infinite' cannot distinguish in which order the limits are taken.

Using limits, the swap is $\lim_{\xi\to\infty}\lim_{y\to\infty}\int_{-\xi}^{+\xi}\int_{-y}^{+y} = \lim_{\xi\to\infty}\lim_{y\to\infty}\int_{-y}^{+y}\int_{-\xi}^{+\xi}$ instead of the *false* $\lim_{\xi\to\infty}\lim_{y\to\infty}\int_{-\xi}^{+\xi}\int_{-y}^{+y} = \lim_{y\to\infty}\int_{-y}^{+y}\lim_{\xi\to\infty}\int_{-\xi}^{+\xi}$.

Because of these transformations, the proof now reduces to showing what he claimed in §235 already, that

> the function represented by the definite integral $\int_0^\infty \cos(\xi(x-\alpha))\,\mathrm{d}\xi$ has the singular property, that if we multiply it by any function $\phi(a)$ and by $\mathrm{d}a$, and integrate it with respect to α between infinite limits, the result is equal to $\pi\phi(x)$; so that the effect of the integration is to change α into x, and to multiply by the number π.

§361

To this end, we first turn our attention to $\int_0^\infty \frac{\sin(x)}{x}\,\mathrm{d}x$. Fourier describes the form of the curve $\frac{\sin(x)}{x}\,\mathrm{d}x$:

> its first ordinate at the origin is 1, and the succeeding ordinates become alternately positive or negative; the curve cuts the axis at the points where $x = \pi, 2\pi, 3\pi \ldots$, and it approaches nearer to this axis.

§415. Observe how he talks about the 'first ordinate'.

The integral is the (signed) area contained by this curve and the x-axis, it is known to equal $\pi/2$. Moreover, $\int_0^\infty \frac{\sin(\xi x)}{x}\,\mathrm{d}x$ also equals $\pi/2$ 'whatever positive number ξ may be'. Supposing that ξ increases without limit, i.e. becomes infinite, then he notes that the oscillations of the 'curve whose ordinate is $\frac{sin(\xi x)}{x}$ are infinitely near. Their base is an infinitely small length equal to $\frac{\pi}{\xi}$'. Comparing the positive area resting on one interval

$\frac{\pi}{\xi}$ with the negative area on the successive one (some finite distance X away from the origin) then we see that

> the expression $\frac{\sin(\xi x)}{x}$ of the ordinate, has no sensible variation in the double interval $\frac{2\pi}{p}$, which serves as the base of the two areas. Consequently the integral is the same as if x were a constant quantity. It follows that the sum of the two areas which succeed each other is nothing.

As Hamilton pointed out, the difference in areas between successive lobes is not merely infinitesimal, but a second order infinitesimal, since the height difference is infinitesimal and their common base is infinitesimal. If it were merely infinitesimal then adding infinitely many of them might still result in a finite discrepancy. A precise argument should have stated this explicitly. (A modern reader might appreciate the same reasoning in the following form: across one interval $\frac{\pi}{\xi}$ the envelope $\frac{1}{x}$ varies by at most $\frac{\pi}{\xi}\cdot\frac{1}{x^2}$, so each pair of lobes leaves a discrepancy in area of at most $\frac{\pi}{\xi}\cdot\frac{\pi}{\xi x^2}$. Since these accumulate to at most $\int_X^\infty \frac{\xi}{\pi}\cdot\frac{\pi^2}{\xi^2 x^2}\,\mathrm{d}x = \frac{\pi}{X\xi}$, the discrepancy vanishes as ξ grows.) He continues: Hamilton 1843, p. 46

> The same is not the case when the value of x is infinitely small, since the interval $\frac{2\pi}{\xi}$ has in this case a finite *ratio* to the value of x. We know from this that the integral $\int_0^\infty \frac{\sin(\xi x)}{x}\,\mathrm{d}x$ in which we suppose ξ to be an infinite number, is wholly formed out of the sum of its first terms which correspond to extremely small values of x. … We express this result by writing
>
> $$\int_0^\infty \frac{\sin(\xi x)}{x}\,\mathrm{d}x = \int_0^\omega \frac{\sin(\xi x)}{x}\,\mathrm{d}x = \frac{1}{2}\pi. \tag{22}$$
>
> §415

(Supplementing our earlier modern argument, it should be noted that we have taken both $\xi \to \infty$ and $X \to \omega$. Since we can always decompose $(0,\infty)$ into $(0,X)$ and (X,∞), we can take $X \to \omega$ after taking $\xi \to \infty$ and the argument is valid. As noted earlier, Fourier's usage of infinitesimals makes him unable to express this difference.)

This granted, §416 considers

$$\frac{2}{\pi}\int_{-\infty}^{+\infty} f(y)\frac{\sin(\xi(y-x))}{y-x}\,\mathrm{d}y \tag{23}$$

Fourier describes a geometric construction: lay down the axes, construct the curve f whose form is 'entirely arbitrary'; construct $\sin(\xi z)/z$, 'z denoting the abscissa and ξ a very great positive number'. This latter curve may be placed wherever and we consider its central peak to be placed at $y-x$. We then 'link together' the two curves to form a third by multiplying each ordinate. The area contained by this third curve and the x-axis is Eq. (23). Now comes the crucial step: for all points which are at a finite distance from the point x

> the value of $f(\alpha)$ varies infinitely little when we increase the distance by a quantity less than $\frac{2\pi}{\xi}$. The same is the case with the denominator $\alpha - x$, which measures that distance. The area which corresponds to the interval $\frac{2\pi}{\xi}$ is therefore the same as if the quantities $f(\alpha)$ and $\alpha - x$ were not variables. §416

But if $f(\alpha)$ is not variable then by the above argument, on all oscillations, the positive and negative part of the oscillation cancel, except for the interval corresponding to $y = x$, where $f(y)$ is considered constant. Hence the area—i.e. Eq. (23)—just equals $f(x)$ since the two factors $\pi/2$ cancel. This concludes Fourier's proof.

The last step involves two premises: Fourier first asserts that $f(x)$ varies infinitely little on an infinitesimal interval $\frac{2\pi}{\xi}$ (on which $\frac{\sin(\xi x)}{x}$ oscillates exactly once); this then leads to the premise that $f(\alpha)$ can be considered constant on that interval. The first of these premises is the modern notion of continuity, which seems to contradict his explicit statement that f is 'entirely arbitrary' with no common 'law' between §417
the values of f. However, would Fourier have considered this a new assumption he tacitly brought in, or would it have been a property that any 'arbitrary' function had? As the geometrical construction makes clear, Fourier takes a function to describe a curve—its values are the ordinates of a curve—and he held that every curve has a finite arclength. A curve of finite arclength is one of bounded variation: the total amount by which the ordinate can increase and decrease over domain is finite. But that is the property that the argument of §416 needs: bounded variation guarantees that the total defects which accumulate from treating f constant on one oscillation $\frac{2\pi}{\xi}$ stays finite, so the residual area vanishes in the limit. Continuity alone does not guarantee this, a continuous function may oscillate without bound on an arbitrarily small interval. However, in spelling out such distinctions I am doing something Fourier is not, it is precisely the act of spelling such distinctions out that would have give more precision to the proof. All in all, I think that Fourier's underlying conception of 'curve' is doing real work in the proof, but the premises it supplies are not spelled out.

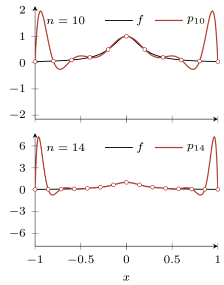
However, apart from the geometric properties of curves, I think that there is a fundamental property of 'function' itself which might have lead Fourier to the idea of considering $f(\alpha)$ to be constant on a single oscillation. The first thing to notice is that considering $f(\alpha)$ to be constant on a single oscillation is the key premise that makes the argument of §415 transfer in §416. If there are other grounds that justify treating $f(\alpha)$ as being constant, then the rest of the proof is valid. In light of this, consider his usage of 'function' discussed in subsection III.3: the arbitrary function f as a sequence of n arbitrary values, these entered into a system of n linear equations involving n coefficients and n functions in which the given function is to be developed. The above proof considers the integral decomposition, but if we consider the analogous case involving series:

$$\frac{1}{2\pi}\int_{-\pi}^{+\pi} f(a) \underbrace{\sum_{n=-\xi}^{\xi} \cos(n(x-\alpha))}_{K_\xi} \mathrm{d}\alpha,$$

then K_ξ describes a curve whose period of oscillation is $\frac{2\pi}{\xi}$. If, like in his treatment of the linear equations, the amount of values defining f is simply ξ, then f is defined by a single value on an interval $\frac{2\pi}{\xi}$—it is *by definition* constant. This perspective comes out of my reinterpretation of what Fourier means by 'function' and is therefore in need of further study. If Fourier's conception of the continuum is indeed, as argued in subsection III.4, that points only exist potentially as divisions of a line, then a function f is only defined with respect to particular resolutions n. These can be arbitrarily refined, but it remains finite at each step. Incidentally, this seems to be how Poisson interpreted the meaning of the series, writing that 'each expansion must be considered as an interpolation formula'.

Poisson 1835, §86

However, following my formalization of the definition in Chapter 3 and without attributing any of that to Fourier, it turns out that *even* if you interpret the statement in such a way, then Fourier's proof is not quite right. Suppose we take f to be given by ξ values and the kernel K_ξ to be defined by ξ terms. A curve fixed at finitely many points may still diverge *between* them as the points are refined. For polynomial interpolation at equally spaced points this is Runge's phenomenon: even a nice function like $1/(1+25x^2)$ on $[-1,1]$ generates an approximating polynomial whose oscillations between the nodes grow without bound. The content of Fourier's theorem, from this perspective, would be that the trigonometric series are *not* like this—that as the number of values n increases, the function (taken as an n-gon) coincides more and more closely with its series rather than oscillating between the prescribed values. The proof in Chapter 3 requires taking *more* terms in the expansion than there are points, i.e. you first let the number of terms in the series go to infinity and only afterwards take the number of points to infinity, not both simultaneously. If you take the limit of the resolution and the number of points together, then the proof requires additional assumptions. However, this involves using a formal definition which Fourier did not have. Moreover, it also involves the kind of order-of-limits distinction that Fourier's usage of infinitesimals had no way to distinguish, and so no way to get right.



Runge's phenomenon for $f(x)=1/(1+25x^2)$: although the interpolants agree with f at every node, increasing the number of equally spaced nodes amplifies the oscillations between them.

The similarity between Fourier's and Dirichlet's proof has generally been recognized by those familiar with Fourier's work. Darboux, the editor of Fourier's *Oeuvres*, described Dirichlet as having followed 'precisément la marche qui est indiquee ici par Fourier, mais en apportant dans les raisonnements une extreme precision' whereas he thought Fourier's proof to be 'exacts dans le fond, bien que manquant de precision'. The afforementioned Gibson expressed the same sentiment in his historical exposition:

Fourier 1888, p. 512

> At bottom Fourier's reasoning is, I believe, quite sound and it seems to me to contain the kernel of Dirichlet's proof. No doubt Fourier did not develope his proof with the extreme precision that the importance of the theorem demanded and that Dirichlet afterwards gave to it; still the substantial accuracy of his reasoning is beyond dispute.

Gibson 1892, p. 137

Contemporary commentators seem to be less generous: Ivor Grattan-Guinness called Fourier's proof a 'fudge', Jesper Lützen puts 'proof' in scare quotes when he refers to it and Clifford Truesdell calls it a 'preposterous legerdemain'. All agree that there is something incomplete about Fourier's proof and that it lacks precision. It is then tempting to jump to the conclusion that it was Dirichlet who supplied the precision which Fourier's work lacked. However, this is what I want to question. I agree that Fourier's proof leaves something to be desired, especially in

Grattan-Guinness 1990, p. 779, Jahnke 2003, p. 157, Truesdell 1980, p. 77.

light of the fundamental importance of the theorem. Most importantly, the decomposition into series other than the trigonometric functions with integer frequencies is incomplete. The geometric construction of Fourier's proof must surely have been, in part, motivated by his awareness that, using Hamilton's words, the 'interior mechanism' which the proof reveals applies more generally. However, the corresponding kernel admits no closed-form expression in the two generalizations he considers: the non-integer frequencies fixed by a transcendental equation, and the decomposition into the Bessel functions of the cylindrical case. In §285 he shows, with a graphical construction, that the roots of the transcendental equation are asymptotically spaced like the integers, and the analogous spacing for the roots of the Bessel functions would not have escaped him. These are the ingredients that could generalize the proof. However, they are only the ingredients, not the proof itself. Its lack is a gap, though a gap in generality and not precision. With respect to imprecision, I described two places in this subsection where Fourier leaves a premise unstated. The first—the second-order cancellation of the area—is a minor omission in an otherwise sound argument. The second, however, is substantial: how one justifies that f can be considered constant across a single oscillation goes to the very question of what 'function' means and what its referents are. If Fourier's 'function' is indeed fundamentally different from Dirichlet's, then we should take seriously the possibility that 'more precision' need not mean 'what Dirichlet did'; if Fourier's 'function' is indeed more like a sequence—a succession of values at finite resolution, not a map on a completed continuum of points—then the missing precision of his work lies in the order-of-limits distinction that his usage of infinitesimals cannot express, and in the lack of a detailed articulation of his own function-concept: a definition as explicit as the one later given for the real numbers. The developments of modern analysis took Dirichlet's definition as basic, and it is therefore easy to take it for granted. The construction of Chapter 3 is intended to challenge this by showing that how Fourier uses 'function' admits a rigorous treatment of its own. It reveals how his use of infinity and infinitesimals is inadequate.

Poisson pointed this out in Poisson 1835, §88.

Hamilton 1843, p.46

The relevant quote is given on p. 51.

For other dubious reasoning see §420–1 and the aforementioned $1-1+1-\ldots=1/2$ of §218. Both of these were pointed out by Gibson 1892.

Chapter 3

A modern formalization of Fourier's 'function'

THE GOAL of the following chapter is to give a definition of 'function' based on Fourier's work that implies the properties that Fourier gives to it; in particular, in which Fourier's proof of the Fourier series representation of arbitrary functions can be expanded into a proof satisfying modern standards of rigor. While all that follows takes inspiration from Fourier's work, none of it should be considered a historical exposition of his ideas. The development below is to be judged purely as mathematics: every claim requires rigorous justification, and the definitions are motivated on mathematical rather than historical grounds. A historical exposition of Fourier's work is given in Chapter 2 subsection III.4.

I will call a *map* any rule which relates a *finite* domain A and some codomain B, i.e. $m : A \to B$ so that to each $a \in A$ there corresponds a *unique* $m(a) = b \in B$. 'Function' can be used in this more generic sense and is done so throughout mathematics. However, that is clearly *not* the concept of function which lies at the foundation of analysis. I use 'map' here to denote something with a finite domain; the crux of the issue of defining 'function' is extending such a concept to *continuous* domains. Conventionally, this is done by first constructing the real numbers and then using the real numbers as the domain of an arbitrary function f, i.e. $f : \mathbb{R} \to \mathbb{R}$.

The definition given below takes its main technical inspiration from the base-a representation of real numbers. A sequence q_n of rational numbers

$$q_n := \sum_{i=-n}^{n} \frac{d_i}{a^i}, \qquad d_i \in \{0, 1, \ldots, a-1\}$$

defines a real number; we will attempt to define functions analogously. The property of the sequence q_n which makes such a definition work is *scale-compatibility*: each element d_i is a unique bit of information in fixed relation to the others: units, tenths, hundredths, thousandths, etc. Increasing the resolution n will add more elements d_i to the sum defining q_n but it does not change the previously identified elements. Measuring in centimeters rather than meters does not change the amount of meters one has. A more abstract definition of the real numbers defines them in terms of equivalence classes of *Cauchy sequences*. This addresses the fact that the same real number can be measured in very different forms (e.g. different bases, continued fraction representation). The Cauchy property is a more general property which implies scale-compatibility: if x_n is a Cauchy sequence, then for any resolution ε one can identify an N beyond which subsequent information refines and does not contradict that of x_N, i.e. $|x_N - x_m| < \varepsilon$ for all $m > N$. Most crucially, a real number is *not* something to which a sequence of rationals tends, it *is* the sequence of rationals; the equality of two real numbers consists in a certain *equivalence* between two sequences of rationals.

Discrete sequences of values are for 'function' what rational numbers are for real numbers (see Fig. 1). The idea will thus be to define a function as a *sequence of maps*, subject to a regularity condition analogous to the Cauchy criterion. When one increases the resolution on a numerical measurement, one compares two quantities and one asserts a bound on their difference. For functions, upon increasing the resolution, multiple refined elements together correspond to a single coarser measurement. In Fig. 1, f^1 of the first row corresponds to f^1 and f^2 on the second row; f^1, f^2, f^3, f^4 on the third row, etc. Now, if $(f^i)_n$ were the average of its m-th descendant generation $(f^j)_{n+m}, (f^{j+1})_{n+m} \ldots$, then additional measurement *refines* but does not *contradict* coarser measurements—we would have scale-compatibility. This means that we can colloquially speak of $(f^i)_\infty$ as a function considered on a 'sufficiently refined scale' analogous to how $x \pm \varepsilon$ with x rational describes a real number on a 'sufficiently refined scale' (and equality between two numbers follows if an arbitrary ε bounds their difference).

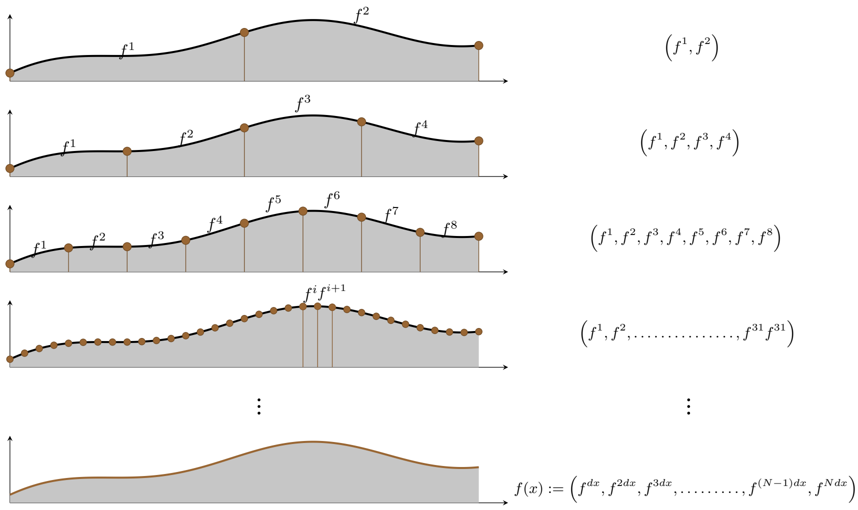


FIGURE 1. The idea of a function as a sequence of maps is the idea that the bottom row refers to the sequence itself, not to something the sequence is continually approaching but never reaching.

I Definition

Consider a finite interval I and $2 \leq a \in \mathbb{N}$. An a-adic partition at resolution n defines a^n uniform parts. These parts, for any resolution, can be labelled by the sequence $(0/a^n, 1/a^n, \ldots (a^n-1)/a^n)$ where each element of the sequence corresponds to a part of I at resolution n. When writing $f_n(x)$, we let x always denote i/a^n for some $0 \leq i \leq a^n - 1$, i.e. an element of $X_n := \{0/a^n, 1/a^n, \ldots, (a^n-1)/a^n\}$.

The use of different partition schemes is discussed below in remark (4).

Definition 3.1. A $\mathcal{F}$unction $f(x)$ is a sequence $(f_n)_{n\in\mathbb{N}}$ of maps

$$f_n : \left\{\frac{0}{a^n}, \frac{1}{a^n}, \ldots, \frac{a^n-1}{a^n}\right\} \to \mathbb{R}$$

subject to the condition that $\lim_{n\to\infty} \varepsilon_n(f) = 0$, where

$$\varepsilon_n(f) := \max_{x\in X_n} \sup_{m\geq 1} \left| \underbrace{f_n(x)}_{\text{Parent}} - \underbrace{\frac{1}{a^m} \sum_{j=0}^{a^m-1} f_{n+m}\left(x + \frac{j}{a^{n+m}}\right)}_{\text{Descendant mean}} \right|. \tag{1}$$

The condition is an asymptotic Martingale condition on an a-ary refinement tree.

In set-theoretic notation:

$$\mathcal{F} := \left\{ (f_n)_{n\in\mathbb{N}} \;\middle|\; f_n : \{0/a^n, 1/a^n, \ldots, (a^n-1)/a^n\} \to \mathbb{R}, \lim_{n\to\infty} \varepsilon_n(f) = 0 \right\} \Big/ \sim .$$

The equivalence relationship '$\sim$' is defined in remark (1).

Some remarks are warranted.

(1) Properly, a $\mathcal{F}$unction is an *equivalence class* of $(f_n)_{n\in\mathbb{N}}$ under the equivalence relationship $(f_n)_{n\in\mathbb{N}} \sim (g_n)_{n\in\mathbb{N}}$ when

$$\forall \varepsilon > 0 \; \exists N \text{ such that } \forall x \in X_n \text{ with } n \geq N \; |f_n(x) - g_n(x)| < \varepsilon. \tag{2}$$

We adopt the standard abuse of writing $f(x) = g(x)$ for $(f_n)_{n\in\mathbb{N}} \sim (g_n)_{n\in\mathbb{N}}$.

Eq. (2) draws out how $\mathcal{F}$unctions differ from conventional functions. For conventional $f, g : \mathbb{R} \to \mathbb{R}$ functions f and g are equivalent if $\forall \varepsilon > 0 \; \forall x \in \mathbb{R} \; |f(x) - g(x)| < \varepsilon$. Thus one assumes 'infinite resolution' on the domain while some-but-any (arbitrarily small) resolution ε suffices on the codomain and it admits functions like $f(x) = \mathbb{1}_{\mathbb{Q}}(x)$. This cannot be defined as a $\mathcal{F}$unction since $\mathcal{F}$unctions only admit specifying some-but-any resolution on the domain and no finite resolution distinguishes rational and irrational numbers. See remark (3) below for a condition under which the value of a $\mathcal{F}$unction can be defined for real number inputs.

$\mathcal{F}$unctions are defined as equivalence classes, motivated by the idea that different sequences of polygons can be used to approximate the same curve, as long as the difference matters less and less as the partition refines. However, each $\mathcal{F}$unction also has a *canonical representative* i.e. the representative f_n for which $\varepsilon_n(f) = 0$ for all n. The canonical representative will be key for defining and proving properties of $\mathcal{F}$unctions.

Lemma 3.1. Each $\mathcal{F}$unction admits a unique *canonical representative* $(f_n)_{n\in\mathbb{N}}$ with $\varepsilon_n(f) = 0$ for all n. Moreover, the canonical representative can be constructed directly from any other representative.

The construction proceeds via cell averages and is therefore deferred to section II, after the integral has been defined.

The canonical representative is precisely the resolution-n data that all representatives of a class have in common. It is therefore the correct object for a pointwise comparison of $\mathcal{F}$unctions. We thus define the pointwise limit of a sequence of $\mathcal{F}$unctions implied by Eq. (2). If $(f^{(m)})_{m\in\mathbb{N}}$ is a sequence of $\mathcal{F}$unctions and g is a $\mathcal{F}$unction all given by their canonical representatives, then $\lim_{m\to\infty} f^{(m)}(x) = g(x)$ if

$$\forall \varepsilon > 0 \quad \forall n \in \mathbb{N}\ \forall x \in X_n \qquad \exists M \text{ such that } \forall m \geq M \quad |f_n^{(m)}(x) - g_n(x)| < \varepsilon. \quad (3)$$

Unlike Eq. (2) we do not have to specify a minimal resolution $N \leq n$ on the domain since Eq. (3) uses canonical representatives.

The use of canonical representatives is necessary. If instead we took arbitrary representatives then it would be natural, like in Eq. (2), to impose $n \geq N$ for some minimal N in Eq. (3) (i.e. some minimal resolution on the domain). However, this does not work since now the condition is not invariant under $\sim$. For example, letting $h_n^{(m)}(x) = 1$ for $n \leq m$ and $h_n^{(m)}(x) = 0$ for $n > m$ defines, for each m, a representative of the zero $\mathcal{F}$unction. Yet, at every fixed n we have $h_n^{(m)}(x) = 1$ for all $m \geq n$, so the sequence would fail to converge to 0.

(2) We use some examples of $\mathcal{F}$unctions to elaborate on the definition.

- Any $f : [0,1] \to \mathbb{R}$ defined by a power series is a $\mathcal{F}$unction in an obvious way: for each $x \in X_n$, evaluate the power series at a standard point of the partition (e.g. endpoints or midpoints.) Note that though we *use* the value at points to define the value of the $\mathcal{F}$unction, the value corresponds to the whole *interval*; it is *not* the value at a point. Consider by analogy the statement 'I am 24 years old.' This statement is not equivalent to 'I am 756.864.000 seconds old', even though these amounts are exactly equal. Rather, '24' refers to the whole *interval*: the lower boundary point is used as a representative for the interval. Similarly, points are the *means* by which we define partitions and values of functions, but they are not their referents.

Observe that the domain of the power series is closed. Open domains introduce subtleties (e.g. $x^{-\alpha}$ on $(0,1)$) which we will leave aside here.

I owe this age example to Pat Corvini in a lecture series 'Two, Three, Four and All That'.

This raises the problem of when and how a single point can be used as a representative for an interval. When one makes measurements of some quantity at discrete points (e.g. every second) then what do these measurements tell us, if anything, about the interval between two measurements? Generally, the problem is thus: given that all measurement is finite and discrete, to determine the values which define a $\mathcal{F}$unction. $\mathcal{F}$unctions do not exhibit some of the pathological behaviour that classical real-number functions can do because they assume a *solution* to this problem. $\mathcal{F}$unctions are not the objects to be measured but forms of measuring given objects; 'temperature at some point over time' is not a $\mathcal{F}$unction, a $\mathcal{F}$unction is the *measurement* of this phenomenon. In the case of smooth curves there is essentially no problem since these are affine at a sufficiently small scale. At such a scale the boundary points are honest representatives of the curve on the interval. However, when no point in an interval represents the interval—for phenomena of unbounded variation, fractal or random structure, multi-phase materials—no obvious method exists to define the values of a $\mathcal{F}$unction. Morally, the value corresponding to the interval must be defined as an average, so that the problem reduces to identifying such an average. This is a non-trivial problem, however. Other ideas, like discretizing the range rather than the domain and using measure-theoretic computations, are necessary for measurement in such situations. Note, in this respect, that for any integer $a \geq 2$, the Borel σ-algebra of $\mathbb{R}$ is generated by the collection of a-adic partitions $\{[ka^{-n},(k+1)a^{-n}) : k \in \mathbb{Z}, n \in \mathbb{Z}\}$ and a Lebesgue integrable function can be recovered a.e. from its averages over them (this follows from the Lebesgue differentiation theorem.) Hence, no essential information is lost in translating a Lebesgue integrable function to a $\mathcal{F}$unction.

- Any Riemann integrable function $f : I \to \mathbb{R}$ defines a $\mathcal{F}$unction by defining the values of the $\mathcal{F}$unction via the Riemann integral over parts of the partition. This embedding is not injective. For example the Thomae function, which is Riemann-integrable, becomes the zero $\mathcal{F}$unction since its Riemann integral equals 0 on any partition. The condition of Riemann integrability compares, for a sequence of refining partitions with parts I_i, the 'infinite resolution' of $\sup_{x \in I_i}$ and $\inf_{x \in I_i}$: respectively the upper and lower sums. The regularity condition of $\mathcal{F}$unctions, by contrast, relates the values at one resolution to each further refined resolution.
- The δ distribution is a $\mathcal{F}$unction. Letting $0 < x_0 < 1$ be some real number,

$$f_n(x) = \begin{cases} a^n & \text{if} \quad x \leq x_0 < x + 1/a^n \\ 0 & \text{else} \end{cases}$$

defines the distribution $\delta(x - x_0)$ as a $\mathcal{F}$unction. A $\mathcal{F}$unction can diverge at points and $\mathcal{F}$unctions are thus not merely a subset of $f : [0,1] \to \mathbb{R}$. One can even construct a $\mathcal{F}$unction that diverges at all points, see Eq. (4).

Thomae's function is the function on $[0,1]$ defined by $f(x) = 1/q$ if $x = p/q \in \mathbb{Q}$ in lowest terms, and $f(x) = 0$ if $x \notin \mathbb{Q}$.

(3) Defining the value of $\mathcal{F}$unctions at points $x \in [0,1] \subset \mathbb{R}$—i.e. constructing a classical $f : [0,1] \to \mathbb{R}$ function—requires the $\mathcal{F}$unction to have a certain level of regularity. To see why consider what the $\varepsilon_n(f)$ condition implies. In the case of a dyadic partition scheme, it implies that $|f_n(x) - \frac{1}{2}\left(f_{n+1}(x) + f_{n+1}(x + \frac{1}{2^n})\right)| < \varepsilon$ for sufficiently large n and $x \in X_n$. However, that does not put any bound on $|f_{n+1}(x) - f_{n+1}(x + \frac{1}{2^n})|$. Concretely, consider the following construction:

$$\begin{gathered} (0) \\ (N, -N) \\ (2N, 0, -2N, 0). \end{gathered} \tag{4}$$

This can of course be iterated and subsequently used to define a $\mathcal{F}$unction. Without formally defining the terms here, the broader phenomenon is that of discontinuous functions of unbounded variation. As discussed in Chapter 2 subsection III.4, Fourier seems to have been unaware of the phenomenon of unbounded variation. Accommodating it is part of why I consider this to be a 'modern' definition, not just in form, but also in content.

Turning back to the issue of defining the value of a $\mathcal{F}$unction at a point, the only feasible way to derive a point-value would be to take a sequence $p_n \to x \in \mathbb{R}$ with $p_n \in X_n$ and then consider $f_n(p_n)$. However, since another sequence q_n can differ arbitrarily little from p_n without a bound on $|f_n(p_n) - f_n(q_n)|$, one cannot in general assign a particular value at a point. In other words, the following implication is generally *false*:

$$p_n \to x \text{ and } q_n \to x \qquad \Rightarrow \qquad f_n(p_n) - f_n(q_n) \to 0. \tag{5}$$

For conventional functions, if we consider p_n and q_n to represent all arbitrary sequences converging to x, then Eq. (5) defines continuity of a function at x. We use it as a definition for the continuity of $\mathcal{F}$unctions too. Consequently, the value of a $\mathcal{F}$unction is defined at a point if it is continuous at that point. If the $\mathcal{F}$unction is continuous at each point then we can then define $f(x)$ for $x \in \mathbb{R}$ so that we obtain a function of real numbers $f : [0,1] \to \mathbb{R}$. Piecewise continuous $\mathcal{F}$unctions have an obvious definition as $\mathcal{F}$unctions; turning them into conventional functions simply requires picking a particular boundary at the points of discontinuity. Note that there exist very irregular $\mathcal{F}$unctions, they just cannot be assigned a (resolution-independent) point-value unless the $\mathcal{F}$unction is continuous.

Lemma 3.2 (Cross-scale comparisons at points of continuity)**.** Let f be continuous at x_0 in the sense of Eq. (5) and let $x_{n_0} \in X_{n_0}$ be the scale-n_0 referent of x_0 (i.e. $x_{n_0} \le x_0 < x_{n_0} + a^{-n_0}$). For every $\varepsilon > 0$ there exists n_0 such that

$$|f_n(z) - f_{n_0}(x_{n_0})| \le \varepsilon \qquad \text{for all } n \ge n_0 \text{ and } z \in X_n,\ |z - x_{n_0}| < a^{-n_0}.$$

In particular, the values of f near x_0 are bounded uniformly in n, and $\lim_{n\to\infty} f_n(p_n)$ exists for every sequence $p_n \to x_0$ with $p_n \in X_n$.

Proof. By Eq. (5), there exist $\rho > 0$ and N such that $|f_n(z) - f_n(z')| < \varepsilon$ when $n \ge N$ and $z, z' \in X_n$ are within ρ of x_0. Choose $n_0 \ge N$ with $a^{-n_0} < \rho$ and $\varepsilon_{n_0}(f) < \varepsilon$. For $n = n_0 + m$, every $z \in X_n$ with $|z - x_{n_0}| < a^{-n_0}$ is within ρ of x_0, so the $f_n(z)$ differ pairwise by less than ε. Each is therefore within ε of their mean, which by the $\varepsilon_n(f)$-condition is within $\varepsilon_{n_0}(f) < \varepsilon$ of $f_{n_0}(x_{n_0})$. Thus the values of f are bounded near x_0. Since any two values at scales $\ge n_0$ are within 2ε of $f_{n_0}(x_{n_0})$, the sequence $f_n(p_n)$ is Cauchy and its limit exists. ☐

(4) The choice of a-adic partitions is baked into the definition of a $\mathcal{F}$unction. A more general definition would require an explicit definition of partition schemes and a way of relating $\mathcal{F}$unctions defined with respect to different partition schemes. To make life easier, we will work purely with the a-adic partition scheme. However, the definitions and properties presented in this chapter can be applied to the following definition of a general partition scheme.

Definition 3.2 (Partition scheme)**.** A *partition scheme* on I is a sequence $\mathcal{P} = (P_n)_{n\ge0}$ in which each P_n partitions I into finitely many disjoint parts, subject to

(i) *(refinement)* each part of P_n is divided into parts of P_{n+1};

(ii) *(mesh)* $\max_{J\in P_n} |J| \to 0$ as $n \to \infty$.

The a-adic scheme is the case where each part is split into a equal parts. The common refinement of two partition schemes indicates a way of comparing $\mathcal{F}$unctions that are a priori defined with respect to different partition schemes. Given two schemes $\mathcal{P} = (P_n)$ and $\mathcal{Q} = (Q_n)$, their *common refinement* $\mathcal{R} = (R_n)$ is the scheme whose points of division at scale n are the union of those of P_n and Q_n. Each R_n refines both P_n and Q_n, and is the coarsest scheme that does. A $\mathcal{F}$unction on $\mathcal{P}$ and a $\mathcal{F}$unction on $\mathcal{Q}$ can be compared by considering both as $\mathcal{F}$unctions on $\mathcal{R}$.

(5) Addition and multiplication extend from numbers to $\mathcal{F}$unctions in the obvious way. $\mathcal{F}$unctions can be added in general but not multiplied in general.

$f+g$ is the $\mathcal{F}$unction $(f_n+g_n)_{n\in\mathbb{N}}$, i.e.

$$(f_n+g_n)(x) := f_n(x)+g_n(x) \quad \text{for} \quad x \in X_n.$$

The definition for the product $f \cdot g$ is analogous:

$$(f_n g_n)(x) := f_n(x) g_n(x) \quad \text{for} \quad x \in X_n$$

However, the example $\delta(x)\cdot\delta(x)$ shows that, for some $\mathcal{F}$unctions, this operation does not define a new $\mathcal{F}$unction.

II Integration

Definition 3.3 (Integral)**.** The discrete integral is defined as the number

$$S^{(n)}_{[x_0,x_1]}(f) := \sum_{i=L_n}^{R_n} f_n(i/a^n)\mathrm{d}x_n$$

Let $\mathrm{d}x_n$ denote the number a^{-n}.

for a-adic numbers $0 \le x_0 < x_1 \le 1$ (i.e. numbers of the form k/a^n for integers $k,n \ge 0$) and $L_n = \lceil a^n x_0 \rceil$ and $R_n = \lceil a^n x_1 \rceil - 1$; in other words, the sum over the parts whose left endpoint lies in $[x_0,x_1)$. The limit is the *integral* of f over $[x_0,x_1]$:

$$\int_{x_0}^{x_1} f(x)\,\mathrm{d}x := \lim_{n\to\infty} S^{(n)}_{[x_0,x_1]}(f).$$

The sequence $S^{(n)}_{[x_0,x_1]}(f)$ is Cauchy, since for $n \ge N$ so that $a^N x_0, a^N x_1 \in \mathbb{N}_0$ the endpoints are exact and

$$|S^{(n)}_{[x_0,x_1]}(f) - S^{(n+m)}_{[x_0,x_1]}(f)| \le (x_1 - x_0)\varepsilon_n(f). \tag{6}$$

Hence the integral always exists.

We can extend the integral to arbitrary real-number endpoints $x_0 < x_1$ if these are points of continuity of the $\mathcal{F}$unction (in the sense of Eq. (5)). In that case, there are n_0 and M with $|f_n(z)| \le M$ for $n \ge n_0$ and $z \in X_n$ that are at most a^{-n_0} away from x_0 (Lemma 3.2). Using the estimate Eq. (6), the additional discrepancy near the endpoints is confined to two strips of width less than a^{-n} inside the cells containing x_0 and x_1. Each contributes at most Ma^{-n}, hence

$$\left|S^{(n)}_{[x_0,x_1]}(f) - S^{(n+m)}_{[x_0,x_1]}(f)\right| \le (x_1 - x_0 + a^{-n})\varepsilon_n(f) + 2Ma^{-n},$$

and the integral exists.

Proposition 3.1. The integral is well-defined (i.e. $(f_n) \sim (g_n)$ implies $\int_{x_0}^{x_1} f\,\mathrm{d}x = \int_{x_0}^{x_1} g\,\mathrm{d}x$), it is linear and is additive over adjacent intervals.

Proof. *Well-definedness:* For every n,

$$\left|S^{(n)}_{[x_0,x_1]}(f)-S^{(n)}_{[x_0,x_1]}(g)\right| \le \sum_{i=L_n}^{R_n} \left|f_n(i/a^n)-g_n(i/a^n)\right| a^{-n}.$$

If $f\sim g$, then given $\varepsilon>0$ there is an N with $|f_n(x)-g_n(x)|<\varepsilon$ for $x\in X_n$ and $n\ge N$, so the RHS is $<\varepsilon(x_1-x_0+a^{-n})$. Since ε is arbitrary, we conclude that the integrals are equal. *Linearity:* $(\alpha f+g)_n=\alpha f_n+g_n$ pointwise, hence $S^{(n)}(\alpha f+g)=\alpha S^{(n)}(f)+S^{(n)}(g)$, now take the limit. *Additivity:* The partitions of $[x_0,x_1]$ and $[x_1,x_2]$ tile that of $[x_0,x_2]$. Hence, $S^{(n)}_{[x_0,x_1]}(f)+S^{(n)}_{[x_1,x_2]}(f)=S^{(n)}_{[x_0,x_2]}(f)$. □

Lemma 3.3 (Consistency with Riemann integral)**.** Let $g:[0,1]\to\mathbb{R}$ be a continuous function (in the classical sense). Then

(*i*) $\left(g|_{X_n}\right)_{n\in\mathbb{N}}$ defines a $\mathcal{F}$unction $E(g)$, with $\varepsilon_n(E(g))\le\omega_g(a^{-n})$;

(*ii*) for $x_0<x_1$, the $\mathcal{F}$-integral of $E(g)$ over $[x_0,x_1]$ equals the Riemann integral of g;

(*iii*) $E(gh)=E(g)E(h)$ pointwise for each $x\in X_n$, so that $E(gh)$ is again a $\mathcal{F}$unction and its $\mathcal{F}$-integral is the Riemann integral of the classical product.

$\omega_g(t):=\sup\{|g(x)-g(y)|: x,y\in[0,1], |x-y|\le t\}$. Continuity on the compact $[0,1]$ gives uniform continuity, hence $\omega_g(t)\to 0$.

Proof. *(i)*: $|g(x)-g(x+j/a^{n+m})|\le\omega_g(a^{-n})$ for $x\in X_n$ and $|j/a^{n+m}|<a^{-n}$. Averaging over j,

$$\left|g(x)-\frac{1}{a^m}\sum_{j=0}^{a^m-1} g\Big(x+\frac{j}{a^{n+m}}\Big)\right|\le\omega_g(a^{-n}),$$

uniformly in m and x. Hence $\varepsilon_n(E(g))\le\omega_g(a^{-n})\to 0$ by uniform continuity, and $E(g)$ is a $\mathcal{F}$unction.

(ii): $S^{(n)}_{[x_0,x_1]}(E(g))=\sum_{i=L_n}^{R_n} g(i/a^n)\,a^{-n}$ is a Riemann sum of g with a uniform partition up to endpoint contributions $2\max_{x\in[0,1]}(|g(x)|)a^{-n}$. Comparing cell by cell,

$$\left|g(i/a^n)\,a^{-n}-\int_{i/a^n}^{(i+1)/a^n} g(x)\,\mathrm{d}x\right|\le\omega_g(a^{-n})\,a^{-n},$$

and summing over the (R_n-L_n+1) cells gives

$$\left|S^{(n)}_{[x_0,x_1]}(E(g))-\int_{x_0}^{x_1} g(x)\,\mathrm{d}x\right|\le(x_1-x_0+a^{-n})\omega_g(a^{-n})+2\max_{x\in[0,1]}(|g(x)|)a^{-n}\xrightarrow[n\to\infty]{}0,$$

where the integrals are Riemann integrals. Since the sum in the LHS converges to the $\mathcal{F}$-integral of $E(g)$, the two are equal.

(iii): At every $x\in X_n$, $(gh)(x)=g(x)h(x)$, so $E(gh)_n=E(g)_n\,E(h)_n$ as maps on X_n. The pointwise product of the two embeddings is therefore the sequence $E(gh)$. Since gh is continuous, *(i)* shows it is a $\mathcal{F}$unction and *(ii)* identifies its $\mathcal{F}$-integral with the Riemann integral of gh. □

Let us now prove that given any representative $(g_n)_{n\in\mathbb{N}}$ the canonical representative of that $\mathcal{F}$unction is

$$(f_n)_{n\in\mathbb{N}},\qquad f_n\left(\frac{i}{a^n}\right):=a^n\int_{i/a^n}^{(i+1)/a^n} g(x)\,\mathrm{d}x \tag{7}$$

and that it is unique.

Proof of Lemma 3.1. By the additivity of the integral over adjacent intervals (Proposition 3.1), $\varepsilon_n(f) = 0$ and f therefore defines a $\mathcal{F}$unction. To show that $f \sim g$, take $\varepsilon > 0$ and choose N such that for all $n \geq N$, $\varepsilon_n(g) < \varepsilon$. This means that

$$\sup_{m\geq 1}\left| g_n\left(\frac{i}{a^n}\right) - \frac{1}{a^m}\sum_{j=0}^{a^m-1} g_{n+m}\left(\frac{i}{a^n}+\frac{j}{a^{n+m}}\right)\right| < \varepsilon.$$

We will use $\sup_{m\geq 1}$ as an upper estimate for $\lim_{m\to\infty}$. Thus, denoting $S^{(n+m)}_{cell}(g) = S^{(n+m)}_{[i/a^n,(i+1)/a^n]}(g)$,

$$\begin{aligned}
&\left| g_n\left(\frac{i}{a^n}\right) - f_n\left(\frac{i}{a^n}\right)\right| = \left| g_n\left(\frac{i}{a^n}\right) - a^n \int_{i/a^n}^{(i+1)/a^n} g(x)\,\mathrm{d}x\right| \\
&= \left| g_n\left(\frac{i}{a^n}\right) - a^n \lim_{m\to\infty} S^{(n+m)}_{cell}(g)\right| = \lim_{m\to\infty}\left| g_n\left(\frac{i}{a^n}\right) - a^n S^{(n+m)}_{cell}(g)\right| \\
&\leq \varepsilon_n(g) < \varepsilon
\end{aligned}$$

for an arbitrary ε. We conclude that $f \sim g$. To show that f is unique, suppose $f \sim f^*$ and that $\varepsilon_n(f^*) = 0$ for all n. Since $f \sim f^*$ we know that there exists an N such that for all $n > N$,

$$\max_{x\in X_n} |f_n(x) - f^*_n(x)| < \varepsilon.$$

Any scale-k value is the mean of descendants. Hence, for every $n > \max(N,k)$, expanding the sum and applying the triangle inequality gives:

$$|f_k(i/a^k) - f^*_k(i/a^k)| \leq \frac{1}{a^{n-k}}\sum_{j=0}^{a^{n-k}-1} |f_n(i/a^k + j/a^n) - f^*_n(i/a^k + j/a^n)| < \varepsilon.$$

Hence, $f_k(x) = f^*_k(x)$ for any k, since ε was arbitrary; they are identical. □

We will generally use the notation '*cell*' for $[i/a^n,(i+1)/a^n]$ in case the argument need not refer to a particular i, thus writing: $\int_{cell} = \int_{i/a^n}^{(i+1)/a^n}$ and likewise $S^{(k)}_{cell}(g) = S^{(k)}_{[i/a^n,(i+1)/a^n]}(g)$.

We would like to indicate how $\mathcal{F}$unctions relate to basic concepts in functional analysis, in particular the existence of Banach spaces of $\mathcal{F}$unctions. One issue, however, is that $|f|$ need not be a $\mathcal{F}$unction if f is a $\mathcal{F}$unction. Specifically, $\varepsilon_n(|f|)$ need not converge to zero (e.g. for the $\mathcal{F}$unction constructed via Eq. (4)). However, even if $|f|$ is not a $\mathcal{F}$unction, we will show in Lemma 3.4 that, using the canonical representative, the sum $\int |f(x)|^p\,\mathrm{d}x = \lim_{n\to\infty} S^{(n)}(|f|^p)$ always exists (if we permit its value to be $+\infty$). Hence, we can speak of $\int |f(x)|^p\,\mathrm{d}x$ for any $\mathcal{F}$unction, even if $|f|^p$ is itself not a $\mathcal{F}$unction. We will define the p-norms this way.

Lemma 3.4 (Pointwise Jensen)**.** For any $\mathcal{F}$unction given by its canonical representative f and any cell, $\left|a^n \int_{cell} f\,\mathrm{d}x\right|^p \leq a^n \int_{cell} |f|^p\,\mathrm{d}x$. Moreover, $n \mapsto S^{(n)}_I(|f|^p)$ is nondecreasing so that $\lim_{n\to\infty} S^{(n)}_I(|f|^p) \in [0,\infty]$ exists.

Proof. Since, for $k>n$, $a^n S^{(k)}_{cell}(f)$ is the mean of the a^{k-n} values of f_k of the subcells, we can apply the finite Jensen inequality for the convex $|\cdot|^p$ to get $|a^n S^{(k)}_{cell}(f)|^p \le a^n S^{(k)}_{cell}(|f|^p)$, then let $k\to\infty$, recognizing the monotonicity of the sums. For the monotonicity, $\varepsilon_n(f)=0$ means that $f_n(x)=a^{-1}\sum_{j=0}^{a-1} f_{n+1}(x+j/a^{n+1})$. Applying the same argument and then summing over all $x\in X_n$ multiplied by a^{-n} gives $S^{(n)}_I(|f|^p)\le S^{(n+1)}_I(|f|^p)$. A nondecreasing sequence converges in $[0,\infty]$ and stays below its limit. □

Definition 3.4. For any $\mathcal{F}$unction and its canonical representative f, the p-norms, for $1\le p<\infty$, are defined as

$$\|f\|_p := \Big(\int_I |f(x)|^p \,\mathrm{d}x\Big)^{1/p}$$

and the inner product as

$$\langle f, g\rangle := \int_I f(x)g(x)\,\mathrm{d}x.$$

Note that, like the norms, $\langle f, g\rangle$ is well-defined as a sum under the condition that $\|f\|_2$ and $\|g\|_2$ are finite, even if the product $f(x)g(x)$ is not a $\mathcal{F}$unction. This follows from the polarization identity applied pointwise to the $S^{(n)}$:

$$S^{(n)}_I(2fg) = S^{(n)}((f+g)^2-f^2-g^2) = S^{(n)}_I((f+g)^2) - S^{(n)}_I(f^2) - S^{(n)}_I(g^2).$$

Theorem 3.1. The p-norms induce Banach spaces of $\mathcal{F}$unctions and $p=2$ in particular induces a Hilbert space. We denote these spaces by $\mathcal{F}^p$.

Proof. The proposition reduces to the following three propositions.

(a) $\mathcal{F}^p$ forms a vector space.

(b) The p-norms are norms.

(c) Any Cauchy sequence of $\mathcal{F}$unctions converges in $\mathcal{F}^p$.

(b): We check that $\|f+g\|_p \le \|f\|_p + \|g\|_p$. For any n,

$$S^{(n)}(|f+g|^p)^{1/p} \le S^{(n)}(|f|^p)^{1/p} + S^{(n)}(|g|^p)^{1/p}$$

by Minkowski's inequality for (finite) sums. Since it holds for any n, we can take limits on both sides. That $\|f\|_p = 0 \Rightarrow f=0$ follows from the monotonicity of $n\mapsto S^{(n)}$ on canonical representatives. We thus have closure under addition so that *(a)* is obvious.

(c): Let $f^{(m)}$ be Cauchy in $\|\cdot\|_p$ where we consider the canonical representative for each $f^{(m)}$, i.e. $\varepsilon_n(f^{(m)})=0$ and $f^{(m)}_n(i/a^n) = a^n\int_{cell} f^{(m)}\,\mathrm{d}x$ for all m,n. We first define a candidate f and subsequently prove that it is in fact in $\mathcal{F}^p$ and that $f^{(m)}$ converges to it in $\|\cdot\|_p$.

Candidate. By the cell bound of Lemma 3.4,

$$\begin{aligned}|f^{(m)}_n(x) - f^{(m')}_n(x)| &= \Big|a^n \int_{cell} (f^{(m)} - f^{(m')})\,\mathrm{d}x\Big| \\ &\le \Big(a^n \int_{cell} |f^{(m)} - f^{(m')}|^p\,\mathrm{d}x\Big)^{1/p} \le a^{n/p}\|f^{(m)} - f^{(m')}\|_p,\end{aligned}$$

so at fixed n the sequence of numbers $f_n^{(m)}(x)$ is Cauchy. Set $f_n(i/a^n) := \lim_{m\to\infty} f_n^{(m)}(i/a^n)$ and $f := (f_n)_{n\in\mathbb{N}}$. Since

$$a^{-k}\sum_{j=0}^{a^k-1} f_{n+k}(i/a^n + j/a^{n+k}) = \lim_{m\to\infty} a^{-k}\sum_{j=0}^{a^k-1} f_{n+k}^{(m)}(i/a^n + j/a^{n+k})$$
$$= \lim_{m\to\infty} f_n^{(m)}(i/a^n) = f_n(i/a^n),$$

f_n is a canonical representative, i.e. $\varepsilon_n(f) = 0$.

Convergence. Since f and each $f^{(m)}$ are canonical representatives, so is $f^{(m)} - f$. For fixed n, pulling $\lim_{m'\to\infty}$ outside a finite sum and subsequently applying Lemma 3.4,

$$S_{[0,1]}^{(n)}(|f^{(m)} - f|^p) = \lim_{m'\to\infty} S_{[0,1]}^{(n)}(|f^{(m)} - f^{(m')}|^p) \le \lim_{m'\to\infty} \|f^{(m)} - f^{(m')}\|_p^p,$$

which is uniform in n. Let $\varepsilon > 0$, choose M so that, for $m, m' \ge M$, $\|f^{(m)} - f^{(m')}\|_p < \varepsilon$. Consequently $\|f^{(m)} - f\|_p < \varepsilon$. Thus also $\|f\|_p \le \|f - f^{(m)}\|_p + \|f^{(m)}\|_p \le \varepsilon + \|f^{(m)}\|_p < \infty$. Hence $f \in \mathcal{F}^p$ and $f^{(m)} \to f$.

The inner product $\langle f, g\rangle$ induces the norm $\|\cdot\|_2$, hence $\mathcal{F}^2$ is a Hilbert space. □

There are many $\mathcal{F}$unctions that are not in any $\mathcal{F}^p$. Noteworthy is that $\delta(x - x_0) \in \mathcal{F}^1$, but $\delta(x - x_0) \notin \mathcal{F}^p$ for $p > 1$.

Note that pointwise convergence in the sense of Eq. (3) does *not* imply $\mathcal{F}^p$-convergence. Eq. (3) states that

$$\forall n \in \mathbb{N}\ \forall x \in X_n \qquad \lim_{m\to\infty} f_n^{(m)}(x) - f_n(x) = 0. \tag{8}$$

Recalling that $f^{(m)}$ and f are canonical representatives, so that $f_n(i/a^n) = a^n \int_{i/a^n}^{(i+1)/a^n} f(x)\,\mathrm{d}x$, Eq. (8) is equivalent to

$$\forall n\ \forall 0 \le i < a^n \quad a^n \lim_{m\to\infty} \int_{i/a^n}^{(i+1)/a^n} (f^{(m)}(x) - f(x))\,\mathrm{d}x = 0. \tag{9}$$

On the other hand, $\mathcal{F}^p$-convergence states that

$$\lim_{m\to\infty} \int |f^{(m)}(x) - f(x)|^p\,\mathrm{d}x = 0.$$

One can recognize in Eq. (9) something that is closer to weak convergence in the theory of L^p-spaces. We thus have the opposite implication, stated in the following corollary.

Corollary 3.1.1. Given a sequence of $\mathcal{F}$unctions, $\|\cdot\|_p$-convergence implies pointwise convergence (in the sense of Eq. (3)).

PROOF. $|f_n^{(m)}(x) - f_n(x)| \le a^{n/p}\|f^{(m)} - f\|_p$. □

A concrete example where the two senses of convergence differ is given by $f^{(m)}(x) = a^{m/2}\mathbb{1}_{[i/a^m,(i+1)/a^m)}(x)$, which defines a canonical representative in an obvious way (via cell averages). Say, $i = 0$, then $f_n^{(m)}(0) = a^{n-m/2} \to 0$ at every fixed n, while $\|f^{(m)}\|_2 = 1$.

III Representation

Let us recall that the purpose of this presentation is to develop a rigorous formalization of ‘function’ inspired by Fourier’s conception of ‘function’. The particular result we were interested in investigating is whether any ‘arbitrary function’ has a Fourier series representation. We now turn to this issue. However, to understand if and how the result that Fourier proved will apply to $\mathscr{F}$unctions, I will first quote explicitly a methodological idea that inspired the subsequent account and gives deeper insight into Fourier’s perspective. Fourier, in an unpublished manuscript about the foundations of geometry, is writing about the use of inscribed polygons in the study of curves:

> The figure of one of these polygons approaches all the more that of the curved line as the number of sides is greater. Now one is going to recognize properties which belong to any one of these polygons and one will conclude from it that they also belong to the curve. Fourier n.d.(a), fol.5[1]

$$\begin{tikzcd} P_n \arrow[r, "n\to\infty"'] \arrow[d, "\Phi"', "\text{property of}"] & C \arrow[d, "\Phi"] \\ \Phi(P_n) \arrow[r, "n\to\infty"] & \Phi(C) \end{tikzcd}$$

Fourier: polygon $\to$ curve

$$\begin{tikzcd} f^{[m]} \arrow[r, "m\to\infty"] \arrow[d, "\Phi"', "\text{property of}"] & f \arrow[d, "\Phi"] \\ \Phi(f^{[m]}) \arrow[r, "m\to\infty"] & \Phi(f) \end{tikzcd}$$

$\mathscr{F}$unction: truncation $\to$ $\mathscr{F}$unction

Figure 2. The structural similarity of Fourier’s statement and our approach.

If we think of $(f_n)_{n\in\mathbb{N}}$ as a finer and finer description of some curve, then any particular element f_m describes a polygon. Our approach will be to *use* f_m in the form of a new $\mathscr{F}$unction $f^{[m]}$—the scale-m truncation of f—that corresponds to this polygon. We then investigate properties of $f^{[m]}$ and derive those of f under $m\to\infty$, see Fig. 2 for the correspondence. The procedure is analogous to considering the embedding of the rational numbers within the real numbers and using properties of rational numbers to derive properties of real numbers. We thus make a distinction between the resolution n of *our perspective* on any function and the resolution m of a new object $f^{[m]}$. Of course, $f^{[m]}$ is a much simpler object—interpreted as a classical function, it is Lipschitz continuous—and we expect it to have properties that can be determined easily and generally. The secondary problem then becomes how to derive those of f from those of $f^{[m]}$ and there might be several ways of defining or extracting a property of f from the sequence of properties of $f^{[m]}$. Let us therefore first define $(f_n^{[m]})_{n\in\mathbb{N}}$ more formally. For $n>m$ we linearly interpolate the values of two neighbours at f_m across the children of the left neighbour, thereby describing a polygon with a^m sides.

Definition 3.5. Given a $\mathcal{F}$unction $(f_n)_{n\in\mathbb{N}}$, the *scale-m truncation* of a $\mathcal{F}$unction is the $\mathcal{F}$unction $f^{[m]}$ defined as

$$f_n^{[m]}(x) := \begin{cases} f_n(x) & n \leq m, \\ \left(1 - \frac{j}{a^{n-m}}\right) f_m(y) + \frac{j}{a^{n-m}} f_m(y + \frac{1}{a^m}) & n > m, \end{cases}$$

where $x = y + \frac{j}{a^n} \in X_n$ with $y \in X_m$ and $0 \leq j < a^{n-m}$.

For $y + \frac{1}{a^m} = 1 \notin X_m$, consider $f_m(1) := f_m(0)$. This choice of course anticipates the use of Fourier series representation.

The truncation depends on the particular choice of representative, but the goal is of course to derive representative-independent properties of f from it, analogous to deriving properties of a real number from the elements of the sequences of rational numbers that define them.

Lemma 3.5. Given a $\mathcal{F}$unction f, the $\mathcal{F}$unction $f^{[m]}$ has the following two properties:

(1) $f^{[m]}$ is bounded: there exists an M_m so that for all $n \in \mathbb{N}$ and $x \in X_n$, $|f_n^{[m]}(x)| < M_m$.
(2) $f^{[m]}$ has a Lipschitz constant L_m: there exists an L_m such that for all $n \in \mathbb{N}$ and $x, y \in X_n$, $|f_n^{[m]}(x) - f_n^{[m]}(y)| < L_m|x - y|$.

Proof. That $f^{[m]}$ is a $\mathcal{F}$unction follows from the inequality $\varepsilon_n(f^{[m]}) \leq L_m a^{-n}$ for $m \leq n$. The two properties are obvious from the definition: for $n \leq m$: $f_n^{[m]}$ is defined by finitely many discrete maps; for $n > m$: $f_n^{[m]}(x)$ is bounded by $\max_{x \in X_m} |f_m(x)|$ and the linear interpolation yields the Lipschitz constant. □

Now we turn to the actual issue of representing $\mathcal{F}$unctions by a Fourier series. The domain of a $\mathcal{F}$unction is an interval I. Until now we have used the real numbers $[0,1]$ to define points in this interval (see Definition 3.3). In the following section it will be more useful to use the numbers $[-\pi, \pi]$. This merely adds a factor 2π in various places.

Fourier was led to the suspicion that it might be possible to represent arbitrary functions with Fourier series from the coefficient formula

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{+\pi} f(x) e^{-ikx} \, \mathrm{d}x \tag{10}$$

Fourier did not use the coefficient formula in the form of complex exponentials, but that is beside the point here. We will consider all statements about complex numbers to be two statements about the real and imaginary part, so that the theory—developed for real numbers—applies.

since this formula is completely general. Fourier considered the product of two functions and the integral of this product to be defined generally; however, since the product of two $\mathcal{F}$unctions is not defined in general, Eq. (10) need not be defined for arbitrary $\mathcal{F}$unctions.

Fourier of course did not have a formalized definition of $\mathcal{F}$unction in the way we developed. Translated to our framework, he simply reasoned with $f_n(x)$ for n 'infinite'. In our framework this means reasoning with the $f^{[n]}$ and then concluding from that properties of f. Retaining the resolution of f in the definition of the coefficients by using the scale-m truncation gives

$$c_k^{[m]} := \frac{1}{2\pi} \int_{-\pi}^{+\pi} f^{[m]}(x) e^{-ikx} \, \mathrm{d}x. \tag{11}$$

If we start here, then it becomes a secondary issue whether Eq. (10) exists and whether $\lim_{m\to\infty} c_k^{[m]} = c_k$, i.e. if we can commute the limit and the integral.

Accordingly, Fourier's main proof for the existence of a Fourier series representation can be interpreted as showing, for a particular n:

$$f_n(x) = \lim_{K\to\infty} \sum_{|k|\le K} c_k^{[n]} e^{\mathrm{i}kx}, \tag{12}$$

and this is the result we will prove below.

This is, however, not a result about a $\mathcal{F}$unction, which is a *sequence* of such f_n. Later Fourier lets n increase together with K i.e. taking as many points in the domain as one takes frequencies. We can interpret as saying that

$$\lim_{K\to\infty} f_K(x) - \sum_{|k|\le K} c_k^{[K]} e^{\mathrm{i}kx} = 0$$

This would of course require $\lim_{K\to\infty} c_k^{[K]}$ to exist and, in our framework, requires additional conditions on f. There is no recognition in Fourier that these two are not equivalent. Both of these two statements are distinct from the statement closest to the classical interpretation of 'any function has a Fourier series representation', namely:

$$f(x) = \lim_{K\to\infty} \sum_{|k|\le K} \left(\lim_{n\to\infty} c_k^{[n]}\right) e^{\mathrm{i}kx}.$$

The three distinct statements are distinguished by how we quantify over the resolution on the domain.

Before we turn to the proof of Eq. (12), one word of clarification. The 'x' in the LHS of Eq. (12) is the argument for a map with a discrete domain where $x = i/a^n$ is first and foremost the label for the interval $[i/a^n, (i+1)/a^n]$. However, the RHS is simply an analytic formula where x should be considered a real number. This is no issue if we simply use the value i/a^n as this real number, but we considered it worth flagging.

Theorem 3.2. Given a representative $(f_n)_{n\in\mathbb{N}}$ of a $\mathcal{F}$unction,

$$f_n(x) = \lim_{K\to\infty} \sum_{|k|\le K} c_k^{[n]} e^{\mathrm{i}kx}, \qquad x \in 2\pi X_n \tag{13}$$

$2\pi X_n := \{-\pi + 2\pi y \mid y \in X_n\}$.

Proof. We substitute Eq. (11) into the RHS of Eq. (13) to obtain

$$\lim_{K\to\infty} \sum_{|k|\le K} \left[\frac{1}{2\pi}\int_{-\pi}^{+\pi} f^{[n]}(y) e^{-\mathrm{i}ky}\,\mathrm{d}y\right] e^{\mathrm{i}kx}$$
$$= \lim_{K\to\infty} \frac{1}{2\pi}\int_{-\pi}^{+\pi} f^{[n]}(y) \sum_{|k|\le K} e^{\mathrm{i}k(x-y)}\,\mathrm{d}y \tag{14}$$

In the proof we use elementary properties of the integral: translation invariance, the estimate $|\int \cdot| \le \int|\cdot|$, that $\int_{-\pi}^{\pi} D_K = 2\pi$, and $\int_{\omega}^{\pi} |\sin(u/2)|^{-1}\,\mathrm{d}u \le 2\log(4/\omega)$. Each could be proven directly for the $\mathcal{F}$-integral, but we do not work out these proofs in the interest of space. The point is that the argument is self-contained and need not refer to the Riemann integral.

Now, interpreted as a classical function of $y \in [-\pi, \pi]$, $f^{[n]}(y)\sum_{|k|\le K} e^{\mathrm{i}k(x-y)}$ is (Lipschitz) continuous. Hence, it also defines a $\mathcal{F}$unction and the two kinds of integrals coincide (by Lemma 3.3). Defining

$$D_K(x-y) := \sum_{|k|\le K} e^{\mathrm{i}k(x-y)} = \frac{\sin((K+\frac{1}{2})(x-y))}{\sin(\frac{x-y}{2})},$$

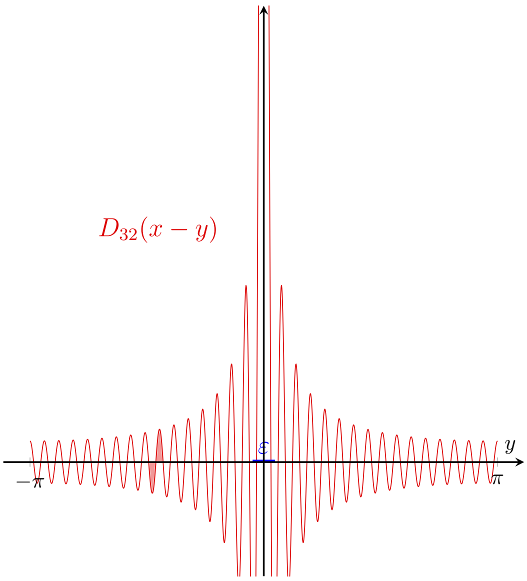


Figure 3. $D_{32}(x-y)$ centered at $y=x$; the shaded areas are equal in the limit.

we first study

$$\int_{-\pi}^{+\pi} D_K(u)\,\mathrm{d}u \tag{15}$$

By a simple computation this equals 2π, independent of K. As K increases, the kernel $D_K(u)$ oscillates more and more rapidly (see Fig. 3), crossing the x-axis every $x_k = k\delta$, where $\delta = \frac{2\pi}{2K+1}$. Define, for each lobe $[x_k, x_{k+1}]$,

$$a_k = \int_{x_k}^{x_{k+1}} D_K(u)\,\mathrm{d}u.$$

Focussing on $[0,\pi]$, we compare, for odd $k>0$, the k-th lobe with the $k+1$-st. Shifting the domain of the second integral by $-\delta$ aligns the bounds of integration and under the shift the lobe changes sign: $\sin((K+\frac{1}{2})(u+\delta)) = -\sin((K+\frac{1}{2})u)$. We combine the integrals into a single integral that measures the increment of the envelope:

$$\begin{aligned}
|a_k + a_{k+1}| &= \left| \int_{x_k}^{x_{k+1}} \frac{\sin((K+\frac{1}{2})u)}{\sin(u/2)}\,\mathrm{d}u + \int_{x_k}^{x_{k+1}} \frac{\sin((K+\frac{1}{2})(u+\delta))}{\sin((u+\delta)/2)}\,\mathrm{d}u \right| \\
&= \left| \int_{x_k}^{x_{k+1}} \sin((K+\tfrac{1}{2})u) \left(\frac{1}{\sin(u/2)} - \frac{1}{\sin((u+\delta)/2)} \right) \mathrm{d}u \right| \\
&\le \int_{x_k}^{x_{k+1}} \left| \frac{1}{\sin(u/2)} - \frac{1}{\sin((u+\delta)/2)} \right| \mathrm{d}u =: |e_k + e_{k+1}| \\
&\le \underbrace{\delta}_{\text{width}} \cdot \underbrace{\left(\frac{1}{\sin(k\delta/2)} - \frac{1}{\sin((k+2)\delta/2)} \right)}_{\text{max height difference}}.
\end{aligned}$$

Where the last step uses the monotonicity of $\sin(u/2)$ for $u \in [0,\pi]$. Adding discrepancies across oscillations—one oscillation being a pair

k and $k+1$—we get a telescoping sum. Consequently, for any fixed $\omega := \frac{2\pi}{2N+1}$ for some $N \in \mathbb{N}$,

$$\left| \int_{\omega}^{\pi} D_K(u)\,\mathrm{d}u \right| \leq \delta \cdot \frac{1}{\sin(\omega/2)} + \delta C_{\omega} \xrightarrow[K\to\infty]{} 0. \tag{16}$$

Here δC_{ω} bounds the sum of *(a)* the last term of the telescoping sum, *(b)* the integral over two partial lobes: from ω to the start of the first paired lobe and the last partial lobe containing π. These three quantities are obviously bounded and $\mathcal{O}(\delta)$. Consequently, all of the mass concentrates at the centre:

$$\frac{1}{2\pi} \int_{-\omega}^{\omega} D_K(u)\,\mathrm{d}u \xrightarrow[K\to\infty]{} 1.$$

We now turn to the general case Eq. (14), repeating the argument for the pure kernel but with an additional factor $f^{[n]}$. We thus choose a resolution n on the domain and a point $x \in 2\pi X_n$.

For notational convenience, $f_m^{[n]}(x)$ is extended 2π-periodically, i.e. $f_m^{[n]}(x+2k\pi) := f_m^{[n]}(x)$ for $x \in X_m$ and $k \in \mathbb{Z}$. This allows us to rewrite, using the evenness of D_K,

$$\begin{aligned}
&\frac{1}{2\pi} \int_{\omega<|x-y|<\pi} f^{[n]}(y) D_K(x-y)\,\mathrm{d}y \\
&\quad = \frac{1}{2\pi} \int_{\omega}^{\pi} f^{[n]}(x-u) D_K(u)\,\mathrm{d}u + \frac{1}{2\pi} \int_{\omega}^{\pi} f^{[n]}(x+u) D_K(u)\,\mathrm{d}u.
\end{aligned} \tag{17}$$

We also define

$$b_k = \frac{1}{2\pi} \int_{x_k}^{x_{k+1}} f^{[n]}(x-u) D_K(u)\,\mathrm{d}u.$$

This leads to

$$\begin{aligned}
2\pi |b_k + b_{k+1}| &\leq \int_{x_k}^{x_{k+1}} \left| \frac{f^{[n]}(x-u)}{\sin(u/2)} - \frac{f^{[n]}(x-u-\delta)}{\sin((u+\delta)/2)} \right| \mathrm{d}u \\
&\leq \int_{x_k}^{x_{k+1}} \left| f^{[n]}(x-u) \left(\frac{1}{\sin(u/2)} - \frac{1}{\sin((u+\delta)/2)} \right) \right| \mathrm{d}u \\
&\quad + \int_{x_k}^{x_{k+1}} \left| \left(f^{[n]}(x-u-\delta) - f^{[n]}(x-u) \right) \frac{1}{\sin(u/2)} \right| \mathrm{d}u \\
&\leq M_n |e_k + e_{k+1}| + L_n \delta \int_{x_k}^{x_{k+1}} \left| \frac{1}{\sin(u/2)} \right| \mathrm{d}u.
\end{aligned}$$

Here the last two estimates are the maximum and Lipschitz constant of $f^{[n]}$ given by Lemma 3.5. We can apply the same argument to the second term of Eq. (17), i.e. for $f^{[n]}(x+u)$. Adding all pairs of lobes except for some arbitrarily small region $[-\omega,\omega]$, gives

$$\frac{1}{2\pi}\left|\int_{\omega<|x-y|<\pi} f^{[n]}(x-u)D_K(u)\,\mathrm{d}u\right|$$
$$\leq \underbrace{\frac{M_n}{2\pi}(\delta\frac{1}{\sin(\omega/2)}+\delta C_\omega)}_{(a)}+\underbrace{\frac{\delta L_n}{2\pi}\int_\omega^\pi\left|\frac{1}{\sin(u/2)}\right|\mathrm{d}u}_{(b)} \quad (18)$$

where (a) is derived using the previous argument. Concerning (b), note that $\int_\omega^\pi\left|\frac{1}{\sin(u/2)}\right|\mathrm{d}u \leq 2\log(4/\omega)$. If $\delta=\frac{2\pi}{2K+1}\to 0$ both (a) and (b) go to zero and what remains is the integral over $[-\omega,+\omega]$.

There we have, noting that $x\in 2\pi X_n$ is fixed,

$$\int_{-\omega}^{+\omega} f^{[n]}(x-u)D_K(u)\,\mathrm{d}u = f_n(x)\int_{-\omega}^{+\omega} D_K(u)\,\mathrm{d}u$$
$$+\int_{-\omega}^{+\omega}\left(f^{[n]}(x-u)-f_n(x)\right)D_K(u)\,\mathrm{d}u$$

Now, for $x\in 2\pi X_n\subset 2\pi X_m$, $f_m^{[n]}(y)=f_n(y)$ and $|f_m^{[n]}(x-u)-f_n(x)|=|f_m^{[n]}(x-u)-f_n^{[n]}(x)|<L_n|u|$ for all $m>n$ so that

$$\left|\int_{-\omega}^{+\omega}\left(f^{[n]}(x-u)-f_n(x)\right)\frac{\sin((K+\frac{1}{2})u)}{\sin(u/2)}\,\mathrm{d}u\right|$$
$$\leq \int_{-\omega}^{+\omega}\left|\frac{f^{[n]}(x-u)-f_n(x)}{\sin(u/2)}\right|\mathrm{d}u \leq \int_{-\omega}^{+\omega}\frac{L_n|u|}{|u|/\pi}\,\mathrm{d}u$$
$$=\int_{-\omega}^{+\omega}\pi L_n\,\mathrm{d}y = 2\omega\pi L_n. \quad (19)$$

Hence, for an arbitrary $\varepsilon>0$ pick $N\in\mathbb{N}$—which determines $\omega=\frac{2\pi}{2N+1}$—so that Eq. (19) is smaller than ε. Then pick $K\in\mathbb{N}$—which determines $\delta=\frac{2\pi}{2K+1}$—so that (a), (b) and $\left|f_n(x)\int_\omega^\pi D_K(u)\,\mathrm{d}u\right|$ are smaller than ε. Combining the estimates of Eqs. (16), (18) and (19), we get

$$\left|\frac{1}{2\pi}\int_{-\pi}^{+\pi} f^{[n]}(y)D_K(x-y)\,\mathrm{d}y - f_n(x)\right| < 4\varepsilon$$

Hence,

$$f_n(x)=\frac{1}{2\pi}\lim_{K\to\infty}\int_{-\pi}^{+\pi} f^{[n]}(y)D_K(x-y)\,\mathrm{d}y$$

and for any $n\in\mathbb{N}$ and $x\in 2\pi X_n$,

$$f_n(x)=\lim_{K\to\infty}\sum_{|k|\leq K} c_k^{[n]}e^{\mathrm{i}kx}.$$

□

Given that $f_n(x)$ is simply a discrete map, a purely finite-dimensional representation of f_n could have been derived using linear algebra by proving that the discrete Fourier transform matrix is invertible. The value of the above result is that, by revealing the 'mechanism' of the result and by using an analytic framework, it suggests paths to extend the result to $\mathcal{F}$unctions, i.e. from concrete $f_n(x)$ to the *sequence* $(f_n)_{n\in\mathbb{N}}$. In particular, if there exists $\omega_K \to 0$ with $\big(M_K/\sin(\omega_K/2) + L_K \log(4/\omega_K)\big)\delta_K \to 0$ and $L_K\omega_K \to 0$, then all upper bounds in the above proof go to zero as $\omega_K \to 0$ so that

$$f_K(x) - \sum_{|k|\le K} c_k^{[K]} e^{\mathrm{i}kx} \xrightarrow[K\to\infty]{} 0. \tag{20}$$

This could be sharpened further by improving the estimates for M_n and L_n since these are currently defined as rough global estimates. For example, if there a finite number of points where f has a jump discontinuity, then we can cut the corresponding pairs of lobes out of the estimates (since $L_n\delta$ does not go to zero there) given that the integral over these cells tends to zero (bounded height and width that converges to zero). However, both of these are statements about particular f_n and not about $\mathcal{F}$unctions as such. All this is to show that, even in a different formalization, the existence of Fourier series representations is a delicate matter where the meaning of the various distinct statements is itself not obvious.

In a similar vein, we had attempted to develop a rigorous account of the derivative for $\mathcal{F}$unctions. However, that too presents various unique difficulties: taking discrete differences on f_n amplifies differences between representatives by a^n and it does not commute with descendant averaging. Given the scope, we decided not to include this account here.

We intend to elaborate on both the Fourier representation of $\mathcal{F}$unctions and the derivative of $\mathcal{F}$unctions in future work.

Chapter 4

The Mellin Transform

THIS CHAPTER develops the Mellin transform $\int_{\mathbb{R}_+} f(x)x^{-s}\frac{\mathrm{d}x}{x}$ and applies it to the asymptotic analysis of elliptic operators on compactified manifolds.

I Two motivating examples from Chapter 1

We begin with two sections with practical and theoretical motivation for the Mellin transform.

I.1 The gravitational potential revisited

Let us recall Laplace's treatment of the gravitational potential in Chapter 1 section I. He wrote down the potential function

$$\int_{\mathbb{R}^3} \frac{f(\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|}\,\mathrm{d}\mathbf{x}' \tag{1}$$

We switch to modern conventions and notation. Laplace's argument required the supporting body be a star-shaped domain. The formulation here, with $f \in C_c^\infty(\mathbb{R}^3)$ supported near the origin and $\Delta : H^2(\mathbb{R}^3) \to L^2(\mathbb{R}^3)$, does not have such restrictions.

and the associated Laplace equation

$$\Delta u = f. \tag{2}$$

To determine u for exterior points (i.e. $r > r'$), he expanded u in powers of $1/r$:

$$u(\mathbf{x}) = \sum_{l=0}^{\infty} \frac{1}{r^{1+l}} a_l Y_l(\theta,\varphi) \tag{3}$$

and computed Y_l in terms of the given data: the shape of the object, which in our case corresponds to the support of f. We had stopped to reflect on the physical meaning of such a decomposition: different properties of the shape—e.g. total mass, center of mass offset, oblateness, etc.—decay at characteristic integral rates. In particular, the associated rate of decay of each Y_l equals $1+l$.

Laplace arrived at this form of the solution by applying a standard mathematical technique: expand a function in powers of one of its variables. But, building on Fourier's spectral insights, we are lead to wonder: are we not decomposing a solution into a product of eigenfunctions just like for the heat equation in a cube, decaying exponentials are eigenfunctions of the temporal part and trigonometric functions are eigenfunctions of the spatial part? The answer is yes, but when we consider the *weighted* Laplacian

$$r^2\Delta = (r\partial_r)^2 + r\partial_r + \Delta_{\mathbb{S}^2}. \tag{4}$$

The theoretical reason for considering the weighted Laplacian will be discussed in the next section, here we will take it for granted and indicate how it gives rise to a spectral perspective. Applying the ansatz $r^s Y(\theta,\varphi)$ to Eq. (4) gives, after dividing by r^s,

$$(s^2+s)Y + \Delta_{\mathbb{S}^2}Y = 0.$$

What we have, then, is not merely a separation of variables $r^s Y(\theta,\varphi)$ but a *decomposition* of the operator: $r^2\Delta$ is built from two operators acting on independent factors—$r\partial_r$ on the radial coordinate, $\Delta_{\mathbb{S}^2}$ on the angular coordinates. The eigenfunctions of such an operator are then products of eigenfunctions of the parts. Laplace's $r^{-(l+1)}Y_l$ ansatz is therefore a consequence: the powers of r are eigenfunctions of the dilation operator $r\partial_r$, the Y_l are eigenfunctions of the spherical Laplacian, and the asymptotic rates $-(l+1)$ are eigenvalues.

This spectral perspective is broader than the concrete computations Laplace used derive his solution. Laplace's computations depend on very particular properties of the problem he considered: (1) the existence of a global radial coordinate r, (2) the integer-valued asymptotic rates and (3) the explicit closed-form of a Green's function Eq. (1). However, these properties do not apply to the analogous problem in the theory of general relativity, fundamental to modern gravitational physics. General relativity replaces Euclidean space with a general Riemannian manifold where no global coordinate system exists, there are no symmetries forcing integer spectra and producing closed-form kernels. Given the astronomical distances involved, we are generally interested in the effects objects have 'far away' i.e. in their asymptotic effects. For Laplace, there was a single coordinate r measuring distance from the source, so that decomposing the potential near the source and infinitely far away come down to the same problem. While there is no canonical radial coordinate r codifying the distance to some object, there is a notion of 'moving infinitely far away' from a given object; while the asymptotic rates are no longer integers, they still form a spectrum; and while there is no explicit Green's function, the concept of a Green's function still applies.

The problem stated generally is this: we seek the asymptotic behavior of solutions to $Pu = f$ on M, where M is a manifold, P an operator on it, and 'asymptotic' refers to a region of M playing the role of $r \to \infty$ in $\mathbb{R}^3$. What would a theory that solves the problem require?

(1) A precise notion of 'going to infinity' compatible with our analytic framework. The solution is to reify 'infinitely far' by attaching a boundary 'at infinity' to the manifold, turning an asymptotic question into a boundary question. *This is compactification.*

(2) A class of operators on this compactified space that behaves, near the boundary, like $r^2\Delta$ did as $r \to \infty$ in $\mathbb{R}^3$. *These are the b-operators.*

(3) A condition on such operators that plays the role of Δ being elliptic and lets us apply the spectral perspective we've been developing. *This is b-ellipticity.*

(4) A way to handle arbitrary data f by decomposing it into modes natural to the geometry at the boundary, and inverting the operator mode by mode. The exponents at which this inversion fails are what generalize Laplace's integer rates. *This is the boundary spectrum.*

The Mellin transform is the analytic tool that makes (4) possible: it will do for our problems what the Fourier transform does for constant-coefficient differential operators, converting the action of a b-operator near the boundary into multiplication by a meromorphic family of operators whose poles are exactly the boundary spectrum.

The purpose of this chapter is to give an account of the Mellin transform. This theory will allow us to prove, in section V, that for an elliptic b-operator P on a compact manifold with boundary, P is Fredholm between weighted b-Sobolev spaces (for weights avoiding the boundary spectrum), and solutions of $Pu = f$ with f smooth admit asymptotic expansions whose exponents are elements of the boundary spectrum.

I.2 Turbulent cascade: multiplicative noise

Let us also recall the urn problem considered in Chapter 1 section II. The setup was combinatorial: to determine the percentage of states with a particular distribution of balls, as a function of an initial distribution. Within the problem, the transition probabilities depended on the *ratio* x/n, not on x and n separately: doubling the number of balls without changing the composition leaves the transition probabilities unchanged. The operator $\mu\partial_\mu$ appears in the resulting PDE

$$\partial_{r'} u = 2u + 2\mu\partial_\mu u + \partial_\mu^2 u. \tag{5}$$

This factor produces a restoring force determined by the *relative* deviation from the mean: doubling μ doubles the pull back to equilibrium. The diffusion ∂_μ^2, by contrast, has the same magnitude everywhere: the noise of an individual swap does not depend on the value of μ. Unlike standard applications of the Laplace transform (or Fourier transform) where a PDE is transformed into an ODE, Laplace's ansatz transformed the PDE into another PDE which required more ingenuity to be solved. There is no single transform that diagonalizes the RHS of Eq. (5). However, to see that a single transform *can* suffice when the dynamics involve only scale-invariant changes, consider the following problem, where the additive randomness of ball-swapping is replaced by a multiplicative randomness compounding across scales.

Problem 4.1 (Random Multiplicative Cascade)**.** Consider a turbulent energy flux cascading from an integral scale L_0. The cascade proceeds in discrete steps $n \in \mathbb{N}$, where at each step the scale is reduced by a fixed factor $\lambda > 1$ (i.e., $L_n = L_0\lambda^{-n}$) and the local dissipation rate ε_n undergoes a random multiplicative update:

$$\varepsilon_{n+1} = W_n \cdot \varepsilon_n \tag{6}$$

where W_n are independent, identically distributed random multipliers with finite moments. We seek the probability density $P_n(\varepsilon)$ of the dissipation rate at the n-th stage of the cascade.

In the continuum limit, the density $P(\varepsilon, t)$ satisfies

$$\frac{\partial P}{\partial t} = \alpha \frac{\partial}{\partial \varepsilon}(\varepsilon P) + D \frac{\partial^2}{\partial \varepsilon^2}(\varepsilon^2 P) \tag{7}$$

where α represents the mean drift and D represents the diffusion of the cascade. Basic algebra transforms Eq. (7) into

$$\partial_t P = [\alpha(\varepsilon \partial_\varepsilon + 1) + D(\varepsilon \partial_\varepsilon + 1)(\varepsilon \partial_\varepsilon + 2)]P.$$

In log coordinates $x = \log \varepsilon$, $p(x) = \varepsilon P(\varepsilon)$, the cascade is a random walk with iid increments $\log W_n$. The diffusion limit gives $\partial_t p = (\alpha + D)\partial_x p + D\partial_x^2 p$, which transforms back into Eq. (7).

Now, using the fact that the Mellin transform $P(\varepsilon, t) \mapsto \hat{P}(s, t)$ diagonalizes $\varepsilon \partial_\varepsilon$ and therefore turns each such term on the RHS into s, we get

$$\partial_t \hat{P}(s,t) = [\alpha(s+1) + D(s+1)(s+2)]\hat{P}(s,t).$$

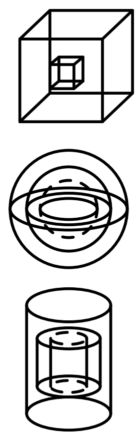
This is just an ODE whose solution is immediate

$$\hat{P}(s,t) = \hat{P}(s,0)\exp([\alpha(s+1) + D(s+1)(s+2)]t). \tag{8}$$

Hence the solution is obtained by applying the inverse Mellin transform to Eq. (8) under appropriate conditions that will be identified in this chapter.

II Why Fourier, why Mellin

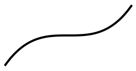
With the foregoing applications in the back of our mind, we can also give a deeper theoretical motivation for the Mellin transform. To this end, let us go back to Fourier's final perspective: the movement of heat in different geometric shapes gives rise to different mathematical decompositions whose effects are observable and are physically significant. Fourier recognized a connection between the different shapes and the resulting decomposition. However, for any particular shape, one can ask a deeper question: why *these* eigenfunctions? Why does the one-dimensional bar lead to trigonometric functions in particular?



The trigonometric functions arose as solutions to an ODE, which was obtained through separation of variables of a PDE. That PDE was itself obtained by examining heat fluxes through infinitesimal volumes of the domain, expressed analytically in terms of infinitesimal changes in the relevant coordinate. It is with the choice of a coordinate system—the choice of independent variables based on the nature of the shape—that the difference begins. For the one-dimensional bar, Fourier considered infinitesimal prisms of volume $l^2\,dx$ and for the sphere infinitesimal shells of volume $4\pi r^2\,dr$. Observe that the volume of the prisms is independent of x while that of the shells depends on r. Consequently, the heat equation in the bar has constant coefficients in x—the equation is the same at every point of the bar—while the radial heat equation has variable coefficients. But sameness amid change is what defines a symmetry, a symmetry *of the equation*. The heat equation on the bar has a translational symmetry, it is invariant under translations $x \mapsto x + a$. The heat equation in the sphere is not invariant under any analogous transformation of r. Additionally, the equation is invariant under a reflection $x \mapsto -x$, but it has no analogous symmetry with respect to time—heat flows from hot to cold.

We now consider the translation-invariance of the heat equation from another angle, by looking for functions which themselves have the same symmetry as the equation. Colloquially, a function f is an eigenfunction of translation if the shape of its graph looks the same everywhere in the domain; mathematically, if $x \mapsto x+a$ has the effect of multiplying $f(x)$ by some scalar, a uniform effect independent of position. Consider the following graphs.

The figure shows $\operatorname{Re} e^{\mathrm{i}x}$.



x^3

$e^{\mathrm{i}x}$

e^{x}

Though the axes have been removed, does the mere graph reveal something about the location of the plotted portions? A graph that did would *not* be translation invariant. In the figure, one is able to compare relative heights of points of the curve. Mathematically, one is able to discern $R_a[f](x) = f(x+a)/f(x)$, but not $f(x)$. Applied to these three, we derive the expressions

$$R_a[x^3] = \left(1+\frac{a}{x}\right)^3 \qquad R_a[e^{\mathrm{i}x}] = e^{\mathrm{i}a} \qquad R_a[e^{x}] = e^{a}.$$

$\cos(x+a) = \cos a \cos x - \sin a \sin x$: translation mixes cos and sin, so neither is an eigenfunction alone. The eigenfunctions of this rotation of $\operatorname{span}\{\cos, \sin\}$ are $e^{\pm \mathrm{i}x}$.

For the exponentials, translation is not a function of location but purely a function of the translation step: a translation $0 \to 1$ has the same effect as $100 \to 101$. Hence, one cannot know if the displayed graph is from $[0,1]$ or $[100,101]$, it looks the same everywhere. We can consider this same idea from another perspective. For which f do we have $T_a f(x) = \tau(a) f(x)$ so that translation by a is an operation independent of the position x and that any plotted part of the curve does not reveal its position? Translations compose, i.e. $T_{a+b} f(x) = T_b T_a f(x)$ which immediately gives the functional equation $f(a+b) = f(a) f(b)$ which is the fingerprint of $e^{\eta x}$, a fact that should not surprise us given the above. Additionally, if η is real, in the case of exponential growth or decay, then there still is an intrinsic directionality, the translations 'to the left' T_a and 'to the right' T_{-a} are inherently different when applied to e^{x}, only $e^{\mathrm{i}\omega x}$ is translation invariant *and* direction invariant.

We thus arrive at a central observation: the translation-invariant functions $e^{\mathrm{i}\omega x}$ are also the eigenfunctions of the translation-invariant heat operator, $\partial_x^2 e^{\mathrm{i}\omega x} = -\omega^2 e^{\mathrm{i}\omega x}$. The same calculation works for any constant-coefficient differential operator—the wave operator, the Schrödinger operator, or any polynomial in ∂_x. Translation-invariant operators act as scalars on $e^{\mathrm{i}\omega x}$. The eigenvalues depend on the particular operator, but the eigenfunctions do not. The following trinity emerges as a summary:

$$\text{translations } T_a \;\leftrightarrow\; \text{generator } \partial_x \;\leftrightarrow\; \text{characters } e^{\mathrm{i}\omega x}.$$

The translation group acts on the bar; its infinitesimal generator is ∂_x; its characters—the functions transforming under translation by a scalar—are $e^{\mathrm{i}\omega x}$. The Fourier transform expresses a function $f(x)$ in terms of its character components, i.e. as $\int_{\mathbb{R}} \hat{f}(\omega)\, e^{\mathrm{i}\omega x}\, \mathrm{d}\omega$. Observe that, with respect

On a finite bar with periodic identification, the translation group is the circle; its characters are the $e^{\mathrm{i}\omega x}$ with $\omega \in \frac{2\pi}{L}\mathbb{Z}$, since a character must return to its value after a complete rotation. The continuum of characters becomes discrete, and the integral a series.

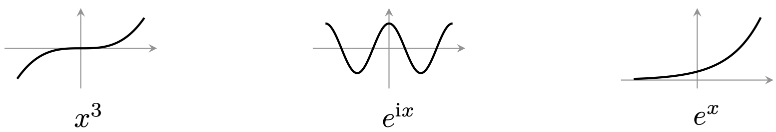
to time, the domain of the heat equation is translation-invariant but intrinsically directional and its eigenfunctions of time are exponential decay terms.

The Mellin transform affords an analogous theoretical motivation. To this end, let us first consider the fact of translation. Translation is a central human activity; we move through space and time and gain awareness of other quantities through such translations: the position of the planets in the sky as time passes, the height of the hill as we move around or the temperature in the room with respect to position and time. Mathematically, we consider other quantities as functions of space and time. The derivative is a measure of the magnitude of change: it measures the change of a quantity in terms of uniform changes in another quantity, just like lengths are measured in terms of uniform units of length—like measures like.

There is, however, another way to measure change: not in terms of translations in the independent variable, but in terms of *dilations* of the independent variable. When we say that gravity is an 'inverse square law', we are making a statement not about the way it changes under a translation in distance, but about how it changes under *dilations* of distance—halving the distance quadruples the force of gravity. When we say an algorithm runs in time $\mathcal{O}(n^2)$, we are making a statement not about how much longer it takes for each additional input, but about how it changes when we scale the input. And when we study a composite material with microstructure at scale ε, we can inquire not only into its properties as we move through the material, but also how they change when we zoom out—when we dilate by $x \mapsto x/\varepsilon$ to pass from the microscopic to the macroscopic scale.

A more comprehensive motivation of the Mellin transform in terms of the dilation group is found in Bertrand et al. 2000.

Let us now recall the three graphs we considered before:



This time, the axes identifying the position of the graphs are there, but no units are given. For which of the graphs can we nevertheless identify a length unit? For which of them can we can recover the unit of length on the x-axis purely from the given equation and graph? The complex exponential obviously reveals one, its period 2π is directly visible. But the real exponential does too: any two heights in a ratio $e : 1$ are separated by a unit of length, which can thus be recovered from the graph alone. To see this systematically, consider the relative sizes we are able to discern: $S_a[f](x) = f(ax)/f(x)$. Applied to our three functions,

$$S_a[x^3] = a^3 \qquad S_a[e^{\mathrm{i}x}] = e^{(a-1)\mathrm{i}x} \qquad S_a[e^x] = e^{(a-1)x}.$$

For the exponentials, S_a retains an x-dependence which allows us to identify the length unit. For the monomial, $S_a[x^3]$ depends only on a: the dilation $1 \to 2$ has the same visual effect as $100 \to 200$, and one cannot tell whether the graph shows $[-2, 2]$ or $[-200, 200]$. The monomial has no intrinsic length scale. The x^s are to dilation what the $e^{\lambda x}$ were to translation: the characters in which dilation acts like a scalar.

To complete the trinity, let us turn back to the weighted Laplacian of section I. Why would we consider $r^2\Delta$ rather than Δ? First, applying Fourier's insight about the physical significance of the flux, note that the Laplacian is the divergence of the gradient. It measures the net flux per unit of volume. However, as we saw earlier, the infinitesimal shells of volume that compose the sphere grow proportionally to r^2. Consequently, Δu combines two effects—the geometric dilution from the expanding shells and the variation of u under dilation and rotation. Multiplying the Laplacian

$$\Delta = \frac{1}{r^2}\partial_r(r^2\partial_r) + \frac{1}{r^2}\Delta_{\mathbb{S}^2}$$

by r^2 cancels precisely the geometric dilution, leaving only the effects of dilation and rotation. The same conclusion follows from dimensional considerations: the asymptotic rate $l+1$ is a dimensionless quantity, but ∂_r has units of inverse length and Δ has two of them: r^2 is exactly the weight needed to make the operator dimensionless and its eigenvalues pure numbers. Concretely: $\Delta r^s = s(s+1)r^{s-2}$ changes the function, whereas $r^2\Delta\, r^s = s(s+1)\, r^s$ extracts the decay rate as an eigenvalue of r^s.

Computing $r^2\Delta$ explicitly gives

$$r^2\Delta = (r\partial_r)^2 + r\partial_r + \Delta_{\mathbb{S}^2}.$$

Observe that this operator contains not ∂_r but $r\partial_r$. Just as ∂_x^2 is a polynomial in the translation generator ∂_x and is therefore diagonalized by the translation characters $e^{\mathrm{i}\omega x}$, so the radial part of $r^2\Delta$ is a polynomial in the dilation generator $r\partial_r$ and is diagonalized by the dilation characters r^s. A second trinity therefore emerges, parallel to the first:

$$\text{dilations } D_a \;\leftrightarrow\; \text{generator } r\partial_r \;\leftrightarrow\; \text{characters } r^s.$$

The Mellin transform is the functional consequence of this second trinity, expressing a function on $\mathbb{R}_+$ in terms of its dilation characters r^s.

The two trinities are linked by a change of coordinates. The substitution $x = e^\xi$ converts dilations of x into translations of ξ, converts $r\partial_r$ into ∂_ξ, and r^s into $e^{s\xi}$. Introducing the function $g(\xi) = f(e^\xi)$ turns dilations of f into translations of g, translating g to the right corresponds to 'zooming out' on f by changing the unit of length. The dilation $x \mapsto rx$ becomes the translation $\xi \mapsto \xi + \log r$ since $f(rx) = g(\log[rx]) = g(\log r + \log x) = g(\xi + \log r)$. Eigenfunctions of translation $g(\xi)$ correspond to eigenfunctions of dilation $f(x)$. This link also extends to their integral transforms. The Fourier transform

$$\mathcal{F}[g](\omega) := \int_{\mathbb{R}} g(\xi)e^{-\mathrm{i}\omega\xi}\,\mathrm{d}\xi, \tag{9}$$

and its scaling analog, the Mellin transform

$$\mathcal{M}[f](s) := \int_{\mathbb{R}_+} f(x)\,x^{-s}\,\frac{\mathrm{d}x}{x} \tag{10}$$

are related via the change of coordinates $x = e^{\xi}$. Writing $s = \sigma + \mathrm{i}\tau$,

$$\mathcal{M}[f](\sigma + \mathrm{i}\tau) = \int_{\mathbb{R}} f(e^{\xi})\, e^{-\xi(\sigma+\mathrm{i}\tau)}\, \mathrm{d}\xi = \mathcal{F}\left[e^{-\sigma\xi} f(e^{\xi})\right](\tau). \tag{11}$$

The Mellin transform of f is the Fourier transform of $e^{-\sigma\xi} f(e^{\xi})$. With σ fixed, the theory of the Fourier transform transfers: inversion, regularity-decay correspondence, Plancherel, convolution. Now consider the role of σ; if $f(x) = x^{\sigma}$, then $e^{-\sigma\xi} f(e^{\xi}) = 1$, its Fourier transform is a delta at $\tau = 0$—we have hit a resonant frequency! The Mellin transform identifies a scaling rate the way the Fourier transform identifies frequency modes. However, treating σ not as a fixed parameter but as variable—letting s range over the complex plane—generally defines a meromorphic function $\mathcal{M}[f](s)$. This comprises the second half of the theory: the Mellin variable s is naturally complex, $\mathcal{M}[f](s)$ is holomorphic on a strip in $\mathbb{C}$ and generally extends meromorphically beyond it, with poles encoding the asymptotic behavior of f.

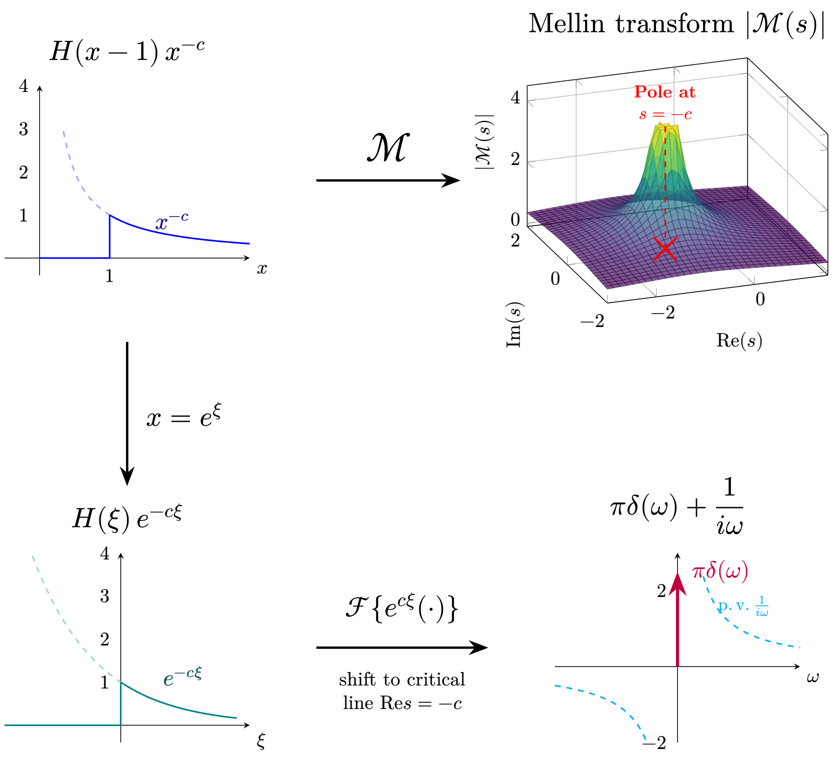


Figure 1. The Mellin transform of $f(x) = H(x-1)x^{-c}$ is meromorphic in s with a simple pole at $s = -c$ (top right). The cutoff $H(x-1)$ removes the singularity. Under $x = e^{\xi}$, f becomes the decaying exponential $H(\xi)e^{-c\xi}$ (bottom left), and evaluating $\mathcal{M}[f]$ along the vertical line through the pole corresponds to taking the Fourier transform of $H(\xi)$ directly resulting in a delta at $\omega = 0$ (bottom right).

III Properties of the Mellin transform

We defer a systematic discussion of the b-Sobolev spaces and b-differential operators to section IV. For now it suffices to introduce weighted L^p spaces.

Definition 4.1. A weighted L^p space is defined as

$$L^{p,\gamma}(\mathbb{R}_+) := \left\{ x^\gamma f \mid f \in L^p\left(\mathbb{R}_+, \frac{\mathrm{d}x}{x}\right)\right\}.$$

III.1 Properties proved via Fourier analysis

This subsection extends the most important properties of the Fourier transform to the Mellin transform. Throughout, we will define $\tilde{f}(\tau) = \mathcal{M}[f](\sigma + \mathrm{i}\tau)$. Crucially, $\mathcal{M}[f]$ is a function of τ only, σ being constant. The proofs consist of a change of variables and an application of the analogous theorem in Fourier analysis, which we take for granted. We will use '$\overset{\mathcal{F}}{=}$' to denote the equality which follows from the related Fourier result and sometimes state it explicitly in the margin.

Definition 4.2. The *inverse Mellin transform* is defined as

$$\mathcal{M}_\sigma^{-1}(\tilde{f})(x) := \frac{1}{2\pi i} \int_{\mathrm{Re}\, s = \sigma} \tilde{f}(s) x^s \, \mathrm{d}s$$

Theorem 4.1 (Mellin inversion)**.** Let $f \in L^{1,\sigma}(\mathbb{R}_+)$ and $\tilde{f} \in L^1(\mathrm{Re}\, s = \sigma)$. Then

$$f(x) = \mathcal{M}_\sigma^{-1}(\tilde{f})(x).$$

Fourier inversion: Let $g, \hat{g} \in L^1(\mathbb{R})$. Then $g(x) = \mathcal{F}^{-1}[\hat{g}(\xi)](x)$.

Proof. Define $g_\sigma(\xi) := e^{-\sigma\xi} f(e^\xi)$, which is equivalent to $f(x) = x^\sigma g_\sigma(\log x)$. Then

$$\tilde{f}(\tau) = \int_{\mathbb{R}_+} f(x) x^{\sigma + \mathrm{i}\tau} \frac{\mathrm{d}x}{x} = \int_{\mathbb{R}_+} g_\sigma(\log(x)) x^{-\mathrm{i}\tau} \frac{\mathrm{d}x}{x} = \int_{\mathbb{R}} g_\sigma(\xi) e^{-\mathrm{i}\tau\xi} \, \mathrm{d}\xi = \hat{g}_\sigma(\tau).$$

Hence,

$$\begin{aligned} f(x) = x^\sigma g_\sigma(\log x) \overset{\mathcal{F}}{=} \frac{x^\sigma}{2\pi} \int_{\mathbb{R}} \hat{g}(\tau) e^{\mathrm{i}\log(x)\tau} \, \mathrm{d}\tau &= \frac{1}{2\pi} \int_{\mathbb{R}} \tilde{f}(\tau) x^{\sigma + \mathrm{i}\tau} \, \mathrm{d}\tau \\ &= \mathcal{M}_\sigma^{-1}(\tilde{f})(x). \end{aligned}$$

□

Proposition 4.1 (Plancherel for Mellin)**.** Let $f \in L^{2,\sigma}(\mathbb{R}_+)$ and define $\tilde{f}(\tau) = \mathcal{M}[f](\sigma + \mathrm{i}\tau)$. Then $\tilde{f} \in L^2(\mathbb{R})$ and

$$\int_{\mathbb{R}_+} |x^{-\sigma} f(x)|^2 \frac{\mathrm{d}x}{x} = \frac{1}{2\pi} \int_{\mathbb{R}} |\tilde{f}(\tau)|^2 \, \mathrm{d}\tau$$

Plancherel for Fourier: Let $f, g \in L^2(\mathbb{R})$. Then $\langle f, g \rangle = \langle \hat{f}, \hat{g} \rangle$.

Proof. Let $g, \hat{g}$ be as in the above.

$$\int_{\mathbb{R}_+} |x^{-\sigma} f(x)|^2 \frac{\mathrm{d}x}{x} = \int_{\mathbb{R}} |g_\sigma(\xi)|^2 \,\mathrm{d}\xi \overset{\mathcal{F}}{=} \frac{1}{2\pi} \int_{\mathbb{R}} |\hat{g}_\sigma(\tau)|^2 \,\mathrm{d}\tau$$
$$= \frac{1}{2\pi} \int_{\mathbb{R}} |\tilde{f}(\tau)|^2 \,\mathrm{d}\tau$$

□

Proposition 4.2 (Regularity-decay for Mellin)**.** Let $f \in L^{1,\sigma}(\mathbb{R}_+)$.

(1) $\mathcal{M}[f](\sigma + \mathrm{i}\tau) \to 0$ as $|\tau| \to \infty$

(2) If $f \in L^{1,\sigma}(\mathbb{R}_+)$ has weak Mellin derivatives up to order k in $L^{1,\sigma}(\mathbb{R}_+)$ then $|\mathcal{M}[f](\sigma + \mathrm{i}\tau)| = \mathcal{O}(|\tau|^{-k})$ as $|\tau| \to \infty$.

Riemann–Lebesgue: If $g \in L^1(\mathbb{R})$, then $\lim_{|\xi|\to\infty} \hat{g}(\xi) = 0$.

Proof. Both follow respectively from (1) the Riemann-Lebesgue lemma and (2) regularity-decay for the Fourier transform. Using the definitions we have used throughout,

$$\tilde{f}(\tau) = \hat{g}_s(\tau) \to 0 \qquad \text{as } \tau \to \infty \qquad \text{by the Riemann-Lebesgue Lemma.}$$

Moreover, f having weak Mellin derivatives up to order k is equivalent to g having weak derivatives up to order k so that

$$|\tilde{f}(\tau)| = |\hat{g}(\tau)| < \mathcal{O}(|\tau|^{-k}) \qquad \text{as} \qquad \tau \to \infty.$$

□

Definition 4.3. The *multiplicative convolution* of f and F is defined as

$$(f \star F)[x] := \int_0^\infty f(y)\, F\!\left(\frac{x}{y}\right) \frac{\mathrm{d}y}{y}.$$

Proposition 4.3 (Convolution for Mellin)**.** Let $f, F \in L^{1,\sigma}(\mathbb{R}_+)$. Then $f \star F \in L^{1,\sigma}(\mathbb{R}_+)$ and

$$\mathcal{M}[(f \star F)](\sigma + \mathrm{i}\tau) = \mathcal{M}[f](\sigma + \mathrm{i}\tau)\mathcal{M}[F](\sigma + \mathrm{i}\tau)$$

Proof. Let $t = e^v \Leftrightarrow v = \log(t)$, so that

$$\begin{aligned}
\mathcal{M}[(f \star F)](\sigma + \mathrm{i}\tau) &= \mathcal{M}\left[\int_0^\infty f(t)\, F\!\left(\frac{x}{t}\right) \frac{\mathrm{d}t}{t}\right](\sigma + \mathrm{i}\tau) \\
&= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} f(e^v) F(e^{u-v}) \,\mathrm{d}v\right] e^{-(\sigma+\mathrm{i}\tau)u} \,\mathrm{d}u \\
&= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} e^{-\sigma u} f(e^v) e^{-\sigma(u-v)} F(e^{u-v}) \,\mathrm{d}v\right] e^{-\mathrm{i}\tau u} \,\mathrm{d}u \\
&= \mathcal{F}[g_\sigma * G_\sigma](\tau) \overset{\mathcal{F}}{=} \hat{g}(\tau)\hat{G}(\tau) \\
&= \tilde{f}(\tau)\tilde{F}(\tau) = \mathcal{M}[f](\sigma + \mathrm{i}\tau)\mathcal{M}[F](\sigma + \mathrm{i}\tau).
\end{aligned}$$

□

III.2 Properties proved via complex analysis

Until now, we have considered σ fixed. However, $\mathcal{M}[f](\sigma + \mathrm{i}\tau)$ becomes more interesting as a complex function of $s = \sigma + \mathrm{i}\tau$. In this subsection, we will use complex analysis to extend the results obtained so far for a single line to regions in the complex plane. By making σ variable, we are considering $L^{2,\sigma}$ spaces with different weights, our purpose being precisely to relate such different spaces. In cases that the weight is not variable, it will therefore be better to use γ to denote the weight.

As a bridge to complex analysis, we first consider a result from Fourier analysis involving an extension into the complex plane: the so-called Paley–Wiener theorem, which has two distinct forms.

Proposition 4.4 (Paley–Wiener for Mellin 1)**.** The following two conditions are equivalent

- $f \in C_c^\infty(\mathbb{R}_+)$ and $\operatorname{supp} f \subset [e^{-R}, e^R]$.
- $\tilde{f}$ is an entire function on $\mathbb{C}$ and

$$|\tilde{f}(s)| \leq C_N (1+|s|)^{-N} e^{R|\operatorname{Re} s|} \qquad \text{for all } s \in \mathbb{C}.$$

C_N denotes a constant that may depend on N.

Proposition 4.5 (Paley–Wiener for Mellin 2)**.** The following two conditions are equivalent

- $f \in L^{2,\gamma}(\mathbb{R}_+)$ and $\operatorname{supp} f \subset [0,1)$.
- $\tilde{f}$ extends to a holomorphic function on $\operatorname{Re} s < \gamma$ and

$$\sup_{\rho<\gamma} \int_{\operatorname{Re} s=\rho} |\tilde{f}(\rho + \mathrm{i}\tau)|^2 \, \mathrm{d}\tau < \infty$$

These two are again a mere change-of-variables from a result of Fourier analysis. However, with respect to the previous section we do get something distinctive: that $\tilde{f}(s)$ extends to an entire or holomorphic function suggests that $\mathcal{M}[f](s)$ should be studied with complex analysis.

Lemma 4.1. Let $\alpha < \beta$ be real numbers. If $f \in L^{2,\alpha}(\mathbb{R}_+) \bigcap L^{2,\beta}(\mathbb{R}_+)$ then

$$f \in \bigcap_{\gamma \in (\alpha,\beta)} L^{2,\gamma}(\mathbb{R}_+)$$

and for $\gamma \in (\alpha, \beta)$

$$\|f\|^2_{L^{2,\gamma}(\mathbb{R}_+)} \leq \|f\|^2_{L^{2,\alpha}(\mathbb{R}_+)} + \|f\|^2_{L^{2,\beta}(\mathbb{R}_+)}.$$

Proof. Let $\gamma \in (\alpha,\beta)$. Define $\theta = \frac{\beta-\gamma}{\beta-\alpha} \in [0,1]$ so that $\gamma = \theta\alpha + (1-\theta)\beta$. Then

$$|f|^2 x^{-2\gamma} = \left(|f|^2 x^{-2\alpha}\right)^\theta \left(|f|^2 x^{-2\beta}\right)^{1-\theta}$$

Integrating and applying Hölder's inequality gives

$$\|f\|^2_{L^{2,\gamma}(\mathbb{R}_+)} \leq \|f\|^{2\theta}_{L^{2,\alpha}(\mathbb{R}_+)} \|f\|^{2(1-\theta)}_{L^{2,\beta}(\mathbb{R}_+)} \leq \|f\|^2_{L^{2,\alpha}(\mathbb{R}_+)} + \|f\|^2_{L^{2,\beta}(\mathbb{R}_+)}$$

□

Proposition 4.6. Suppose $f \in L^{2,\alpha}(\mathbb{R}_+) \bigcap L^{2,\beta}(\mathbb{R}_+)$. Then $\tilde{f}(s)$ is holomorphic in the vertical strip $\alpha \leq \operatorname{Re} s \leq \beta$ and

$$\|\tilde{f}\|_{L^2(\operatorname{Re} z=\gamma)} < \|\tilde{f}\|_{L^2(\operatorname{Re} z=\alpha)} + \|\tilde{f}\|_{L^2(\operatorname{Re} z=\beta)} \tag{12}$$

for all γ with $\alpha < \gamma < \beta$.

Proof. Let $f \in L^{2,\alpha}(\mathbb{R}_+) \bigcap L^{2,\beta}(\mathbb{R}_+)$. Define a smooth cutoff function $\omega(x)$ that is $=1$ near $x=0$ and $=0$ for $x>2$, split $f = f_0 + f_\infty := \omega f + (1-\omega) f$. Concerning f_0, Proposition 4.5 implies that $\tilde{f}_0$ extends to a holomorphic function on $\operatorname{Re} s < \beta$. Concerning f_∞, first note that $\mathcal{M}[f(1/x)](s) = \mathcal{M}[f(x)](-s)$; thus Proposition 4.5 applied to $f_\infty(1/x)$ gives that $\tilde{f}_\infty(s)$ extends to a holomorphic function on $\operatorname{Re} s > \alpha$. We conclude that $\tilde{f} = \tilde{f}_0 + \tilde{f}_\infty$ extends to a holomorphic function in $\operatorname{Re} s \in (\alpha, \beta)$.

Lemma 4.1 together with Plancherel for Mellin (Proposition 4.1) implies Eq. (12). □

Recall that the poles of the Mellin transform correspond to asymptotic modes. This idea will guide our extension of this result in the rest of the section. We will work towards a *global* inversion theorem, i.e. an inversion theorem for $\mathcal{M}[f](s)$ as a complex function corresponding to some f with a particular kind of asymptotic behavior.

To this end, we start with a fundamental result of complex analysis.

Lemma 4.2 (Contour deformation)**.** Let $h(s)$ be holomorphic in a domain D. Let $\Gamma_{0,T}, \Gamma_{1,T}$ be two oriented contours so that for any $T > 0$ there exist arcs A_T, B_T such that $\Gamma_{0,T} + A_T - \Gamma_{1,T} + B_T$ is a closed contour in D. Whenever

$$I_j := \lim_{T\to\infty} \int_{\Gamma_{j,T}} h(s)\,\mathrm{d}s$$

exist and

$$\lim_{T\to\infty} \int_{A_T} h(s)\,\mathrm{d}s = \lim_{T\to\infty} \int_{B_T} h(s)\,\mathrm{d}s = 0.$$

Then

$$I_0 = I_1.$$

This result can be found in any book on complex analysis.

Lemma 4.3 (Contour deformation with poles)**.** Consider the same conditions as in Lemma 4.2, except that h is meromorphic on D and is positively (counterclockwise) oriented. Additionally suppose
- h has no poles on $\Gamma_{0,T} + A_T - \Gamma_{1,T} + B_T$;
- the set of poles of h inside the contour is finite.

Then

$$I_0 = I_1 + 2\pi i \sum_{\rho\in P} \operatorname{Res}(h, \rho),$$

where P is the set of poles of h swept across by deformation.

This is the residue theorem for the closed, positively oriented contour $\Gamma_{0,T} + A_T - \Gamma_{1,T} + B_T$.

Theorem 4.2 (Global Mellin Inversion)**.** Let $f : \mathbb{R}_+ \to \mathbb{C}$ denote a function and, supposing it exists, $\tilde{f}$ its Mellin transform. Whenever,
- $f(x) = \mathcal{M}_\sigma^{-1}[\tilde{f}](x)$ holds for some fixed σ.
- $\tilde{f}(s)$ is meromorphic in a region D that contains the base contour (the line $\operatorname{Re} s = \sigma$), another contour Γ and connecting segments for each truncation.
- The poles P of $\tilde{f}(s)$ are locally finite in any vertical strip.
- Growth bounds on $\tilde{f}$ in D are such that: the integral on Γ converges and the connecting arcs in the deformation vanish.

Then, with all vertical contours oriented upward and Γ lying to the right of the base line $\operatorname{Re} s = \sigma$,

$$f(x) = \frac{1}{2\pi i} \int_\Gamma \tilde{f}(s) x^s \, \mathrm{d}s - \sum_{\rho \in P} \operatorname{Res}\left(\tilde{f}(s)x^s, \rho\right). \tag{13}$$

Proof. Apply Lemma 4.3 with $\Gamma_0 = \Gamma$ and Γ_1 the base line, both oriented upward and Γ to the right. The contour $\Gamma_0 + A - \Gamma_1 + B$ is positively oriented, so $I_\Gamma = I_{\text{base}} + 2\pi i \sum_{\rho \in P} \operatorname{Res}$. Since $f(x) = \frac{1}{2\pi i} I_{\text{base}}$, the stated identity is obtained by rearranging. $\square$

We will state three practical corollaries of this general theorem.

Definition 4.4.

- Let $D \subset \mathbb{C}$, and let $\tilde{f}$ be meromorphic on D. We denote by $\tilde{f}^\star$ the holomorphic part of $\tilde{f}$, i.e. the function obtained by locally subtracting the principal parts of $\tilde{f}$ at its poles.
- For a meromorphic function $\tilde{f}$, we denote by $\mathcal{E}(\tilde{f})$ the index set of poles: the tuples (z,k) with z a pole of $\tilde{f}$ and k an integer satisfying $0 \le k \le \operatorname{ord}_z(\tilde{f}) - 1$ where $\operatorname{ord}_z(\tilde{f})$ is the order of the pole at z. A pole of order m at z thus contributes the m tuples $(z,0), \dots, (z, m-1)$, one for each power of $\log x$ that the residue of $\tilde{f}(s)x^s$ at z produces. We abbreviate $\mathcal{E}_\gamma(\tilde{f}) := \{(z,k) \in \mathcal{E}(\tilde{f}) : \operatorname{Re} z < \gamma\}$.
- We use $S_{\alpha,\beta}$ to denote the vertical strip $\{z \in \mathbb{C} \mid \alpha \le \operatorname{Re} z \le \beta\}$.
- A function $\tilde{f}(z)$ has vertical Schwartz decay on a strip $S_{\alpha,\beta}$ if

$$|\tilde{f}(z)| \le C_N \langle \operatorname{Im} z \rangle^{-N} \quad \text{for all } N.$$

C_N denotes a constant that may depend on N.

Corollary 4.2.1. Assume $f(x) = \mathcal{M}_\alpha^{-1}[\tilde{f}](x)$ and

(i) $\tilde{f}$ is meromorphic on $S_{\alpha,\beta}$;
(ii) the poles of $\tilde{f}$ in $S_{\alpha,\beta}$ are locally finite in vertical strips;
(iii) $\tilde{f}^\star \in L^2(\operatorname{Re} s = \gamma)$ for $\gamma \in \{\alpha, \beta\}$.

Then for every target line $\operatorname{Re} s = \gamma$ with $\gamma \in (\alpha, \beta)$,

$$f(x) = \sum_{(z,k) \in \mathcal{E}_\gamma} a_{z,k} x^z (\log x)^k + r_\gamma(x), \qquad r_\gamma \in L^{2,\gamma}(\mathbb{R}_+),$$

for the index set $\mathcal{E}_\gamma(\tilde{f}) := \{(z,k) \in \mathcal{E}(\tilde{f}) : \operatorname{Re} z < \gamma\}$.

Corollary 4.2.2. Assume $f(x) = \mathcal{M}_\alpha^{-1}[\tilde{f}](x)$ and

(i) $\tilde{f}$ is meromorphic on $S_{\alpha,\beta}$;
(ii) the poles of $\tilde{f}$ in $S_{\alpha,\beta}$ are locally finite in vertical strips;
(iii) $\langle \operatorname{Im} s \rangle^m \tilde{f}^\star \in L^2(\operatorname{Re} s = \gamma)$ for $\gamma \in \{\alpha, \beta\}$.

Then for every target line $\operatorname{Re} s = \gamma$ with $\gamma \in (\alpha, \beta)$,

$$f(x) = \sum_{(z,k) \in \mathcal{E}_\gamma} a_{z,k} x^z (\log x)^k + r_\gamma(x), \qquad r_\gamma \in H_{\mathrm{b}}^{m,\gamma}(\mathbb{R}_+),$$

$H_{\mathrm{b}}^{m,\gamma}$ are the b-Sobolev spaces defined in Definition 4.8.

for the index set $\mathcal{E}_\gamma(\tilde{f}) := \{(z,k) \in \mathcal{E}(\tilde{f}) : \operatorname{Re} z < \gamma\}$.

Corollary 4.2.3. Assume $f(x) = \mathcal{M}_\alpha^{-1}[\tilde{f}](x)$ and

(i) $\tilde{f}$ is meromorphic on $S_{\alpha,\beta}$;
(ii) the poles of $\tilde{f}$ in $S_{\alpha,\beta}$ are locally finite in vertical strips;
(iii) $\tilde{f}^\star$ has vertical Schwartz decay on $S_{\alpha,\beta}$.

Then for every target line $\operatorname{Re} s = \gamma$ with $\gamma \in (\alpha, \beta)$,

$$f(x) = \sum_{(z,k)\in\mathcal{E}_\gamma} a_{z,k}\, x^z (\log x)^k + r_\gamma(x), \qquad r_\gamma \in \mathcal{A}^\gamma,$$

for the index set $\mathcal{E}_\gamma(\tilde f) := \{(z,k) \in \mathcal{E}(\tilde f) : \operatorname{Re} z < \gamma\}$.

$\mathcal{A}^\gamma$ is the space of conormal functions defined in Lemma 4.10.

In the above corollaries, we can allow $\alpha = -\infty$ or $\beta = +\infty$. Also, if $\tilde f$ is holomorphic on $S_{\alpha,\beta}$, then $\mathcal{E}(\tilde f)$ is empty, thus $f(x) = r_\gamma(x)$.

Supposing that $\tilde f$ extends to a meromorphic function on $\mathbb{C}$, we can loosen the conditions on vertical decay of $\tilde f$, i.e. the regularity we demand of f. The following lemma makes this trade-off explicit.

Lemma 4.4. Let Γ be a C^1 curve parameterized by $s(t) = \sigma(t) + i\tau(t)$, $t \in \mathbb{R}$. Suppose in a region containing Γ the growth bound

$$|\tilde f(\sigma + i\tau)| \le A(\sigma) B(|\tau|)$$

holds. Then

$$\int_\Gamma \left|\tilde f(s) x^s\right| \mathrm{d}s \le \int_{\mathbb{R}} A(\sigma(t)) B(|\tau(t)|) x^{\sigma(t)} \, |s'(t)| \, \mathrm{d}t.$$

We will state two concrete trade-offs that might be of value. Assuming $s(t) = \sigma(t) + it$ and that $A(\sigma)$ grows at most exponentially in σ then,

- if $B(|t|) \sim (1+|t|)^m$: choose $\sigma(t) \sim -\kappa \log(1+|t|)$, $\kappa > 0$. We get a bound on the integrand of the form $A(\sigma(t))(1+|t|)^{m-\kappa\log(x)}$.
- If $B(|t|) \sim e^{a|t|}$: choose $\sigma(t) \sim -\kappa|t|$ with $\kappa \log x > a$. We get a bound of the form $A(\sigma(t)) e^{-(\kappa\log(x)-a)|t|}$

These apply to asymptotics as $x \to \infty$. Analogous conditions can be stated for $x \to 0$.

Corollary 4.2.4 (Trading regularity for continuation)**.** Assume

(i) $\tilde f$ is meromorphic in some region D containing $\operatorname{Re} s = \gamma$;

(ii) the poles of $\tilde f$ in D are locally finite;

(iii-a) $\tilde f \in L^2(\operatorname{Re} s = \gamma)$ $\quad (\Leftrightarrow f \in L^{2,\gamma}(\mathbb{R}_+))$;

(iii-b) Some curve Γ in D (as in Lemma 4.4) gives the growth bound

$$|\tilde f(\sigma + \mathrm{i}\tau)| \le A(\sigma) B(|\tau|);$$

(iii-c) the connecting arcs of $\operatorname{Re} s = \gamma$ and Γ give integrals which go to zero (e.g. if the connecting arc Γ_T is parameterized by $\sigma_T(t) + \mathrm{i}\tau_T(t)$, $t \in [0,1]$, and $\operatorname{len}(\Gamma_T) \sup_{t\in[0,1]} \left[A(\sigma_T(t)) B(|\tau_T(t)|) x^{\sigma_T(t)}\right] \to 0$);

then

$$f(x) = \sum_{(z,k)\in\mathcal{E}(\tilde f)} a_{z,k}\, x^z (\log x)^k + \frac{1}{2\pi i} \int_\Gamma \tilde f(s) x^s \, \mathrm{d}s.$$

Proof. Deforming $\operatorname{Re} s = \gamma$ to Γ, the connecting arcs vanish by (iii-c) with Lemma 4.4, leaving the residue sum and $\frac{1}{2\pi i} \int_\Gamma \tilde f(s) x^s \, \mathrm{d}s$. $\square$

These corollaries are worth stating in their own right. By stating various forms, they also reveal an underlying trade-off. The vertical decay of $\tilde f^\star$ corresponds to the regularity of the remainder r_γ, and weakening one coarsens the other: Schwartz decay gives a conormal remainder, polynomial decay a b-Sobolev one, and L^2 membership an L^2 remainder. The same correspondence continues past the cases we have stated: polynomial *growth* in $\operatorname{Im} s$ places r_γ in negative-order b-Sobolev spaces, and in the limit in spaces of distributions. Also, the target line $\mathrm{Re}\, s = \gamma$ fixes the guaranteed order x^γ of the remainder. We do not pursue further generalizations here. One may allow $\tilde{f}$ to be operator-valued, at the cost of more delicate norm bounds in the contour integral. This is the setting the b-calculus application will in fact require. One may also relax the hypothesis that the singularities of $\tilde{f}$ are discrete: where $\tilde{f}$ fails to be holomorphic on a region rather than at isolated poles, the discrete sum $\sum x^z (\log x)^k$ is replaced by a continuum of power-law modes. This is Schulze's framework of *continuous asymptotics*, see p. 124.

IV Compactification and the b-calculus

Let us recall the overarching goal: to study the asymptotic behavior of a quantity u subject to some PDE. If we have an analytic expression for u, like Laplace did for the gravitational potential, then we can deduce the asymptotic behavior of u simply from the analytic expression. However, in the more general setting of differential operators P on manifolds M no such analytic expression exists. The mathematical theory which considers such general operators is functional analysis, in particular the theory of elliptic operators on manifolds. The classical theory proceeds in two steps. Locally, at any point $x_0 \in M$ the principal symbol of P (the leading-order part with coefficients frozen at x_0) controls the operator's fine-scale behavior, which is then analyzed in the frequency domain via the Fourier transform on the tangent space. Ellipticity of the symbol gives local regularity (solutions inherit smoothness from the data) and a local approximate inverse, the parametrix. Globally, when M is compact, these local parametrices patch together to give an inverse modulo compact operators, so that P is Fredholm between appropriate Sobolev spaces. These results do not extend to asymptotic analysis in an obvious manner. In the non-compact setting, there is no 'place at infinity' at which we can freeze the coefficients to define a model characterizing the asymptotic regime. Moreover, the Fredholm property fails on a non-compact manifold and the spectrum is no longer discrete. The standard functional analytic framework applies to *local analysis*—that is, to specific locations within a compact manifold—but does not apply to asymptotic analysis.

Our solution will be to *reify* 'infinity', i.e. to consider 'going to infinity' as a boundary of the manifold, turning the asymptotic analysis of the non-compact space into the analysis of the behavior at a boundary of a compact space. This geometrical context in place, we then turn to the operators describing asymptotic behavior. We saw in section II that $r^2\Delta$ characterizes asymptotic behavior, since the radial modes r^s are its eigenfunctions; Δ does not, since $\Delta r^s = s(s+1)r^{s-2}$. This difference reappears as a degeneracy of Δ at the boundary. Using $\rho = 1/r$, we compute $\Delta = \rho^4 \partial_\rho^2 + \rho^2 \Delta_{\mathbb{S}^2}$. The coefficients vanish at the boundary $\{\rho = 0\}$, and freezing them there gives the zero operator: there is no meaningful model of Δ at the boundary. By contrast, $r^2\Delta = (\rho\partial_\rho)^2 - \rho\partial_\rho + \Delta_{\mathbb{S}^2}$, is a meaningful dilation-invariant model. Dilation-invariance and non-degeneracy at the boundary are two perspectives on the same property.

We therefore identify the class of operators whose degeneracy at the boundary is of a specific kind (the b-operators), define their model at the boundary (the normal operator), and develop the corresponding notion of ellipticity (b-ellipticity). Within this framework (the b-calculus) we then state the analogues of classical Fredholm and regularity results, now between weighted b-Sobolev spaces.

IV.1 The domain and its operators

Spherical coordinates (r,θ,φ) identify Euclidean space with $\mathbb{R}_{>0}\times\mathbb{S}^2$. Its compactification essentially consists of a naive idea: attaching (∞,θ,φ) as a place, a sphere glued to $\mathbb{R}_{>0}\times\mathbb{S}^2$ 'at $r=\infty$.' In other words, the radial compactification $\overline{\mathbb{R}^3}$ of $\mathbb{R}^3$ is diffeomorphic to the closed unit ball $\overline{B^3}$ with boundary $\mathbb{S}^2$ representing $r=\infty$. Moreover, we have the boundary defining function $\rho=1/r$ (extended smoothly across $\rho=0$). The crux is that such constructions are possible for many other manifolds, not just for $\mathbb{R}^3$. There are many kinds of compactification, but the one relevant to the b-calculus essentially consists of constructing a boundary defining function. The specific details depend on the particular manifold and are beyond our scope. However, the subsequent analysis is completely general, it considers a manifold *qua* compact Riemannian manifold with boundary. We work with a compact manifold M with boundary ∂M, equipped with a boundary defining function ρ formally defined below. In local coordinates near the boundary we write $(\rho,y)=(\rho,y_1,\dots,y_{n-1})$, where y_i are coordinates on ∂M. While we will use $\overline{\mathbb{R}^3}$ as our model example, we should keep in mind that the motivation for the heavy machinery is precisely its generality beyond $\mathbb{R}^3$.

We now define the class of b-operators, generalizing the properties of $r^2\Delta$ at the boundary. The operator $r^2\Delta$ is built from $\rho\partial_\rho$, ∂_θ and ∂_φ, which are all tangent to the boundary. They define vector fields which, at $\rho=0$, do not generate motion along the direction of ρ: the angular $\partial_\theta,\partial_\varphi$ generate motion along the boundary while $\rho\partial_\rho$ vanishes at the boundary. Generalizing to operators with variable coefficients, the standard technique consists of constructing a local model by freezing the coefficients. Freezing the coefficient $a(\rho,y)$ at a boundary point $\rho=0$ requires a to be well-defined there, excluding e.g. $a=1/\rho$. In fact, it makes sense to consider a smooth. This affords Taylor expansions of the coefficients at $\rho=0$ and ensures the operator class is closed under composition. These two conditions together specify the relevant class of vector fields: those that are *smooth* on M and *tangent* to ∂M. Operators built from such vector fields define a model operator at the boundary; they have a well-defined model for their asymptotic behavior, considered as their behavior at the boundary of the manifold.

Definition 4.5. For a smooth manifold M with boundary ∂M, we define:

- A *boundary defining function*: a nonnegative function $\rho\in C^\infty(M)$ so that $\partial M=\rho^{-1}(0)$ and $\mathrm{d}\rho(p)\neq 0$ for all $p\in\partial M$.
- The *b-vector fields* $\mathcal{V}_{\mathrm{b}}(M)\subset\mathcal{V}(M)$: the smooth vector fields tangent to ∂M.

Here $\mathrm{d}\rho(p)\in T_p^*M$ is the differential of ρ at p: $\mathrm{d}\rho(p)(v)=v\rho$, $v\in T_pM$. Thus $\mathrm{d}\rho(p)\neq 0$ means that some directional derivative of ρ at p is nonzero.

- The *b-differential operators* $\mathrm{Diff}_{\mathrm{b}}^{m}(M)$: finite sums of products of b-vector fields. $\mathrm{Diff}_{\mathrm{b}}^{0}(M) := C^\infty(M)$

A Riemannian manifold with boundary has a *collar*: a neighbourhood diffeomorphic to $[0,\varepsilon)_\rho \times \partial M$, with ρ corresponding to the boundary defining function. In the collar, any $P \in \mathrm{Diff}_{\mathrm{b}}^{m}(M)$ can be written in the form

$$P = \sum_{j=0}^{m} P_j(\rho)(\rho\partial_\rho)^j \tag{14}$$

with smooth $P_j : [0,\varepsilon)_\rho \to \mathrm{Diff}^{m-j}(\partial M)$.

Lemma 4.5. $\mathcal{V}_{\mathrm{b}}(M)$ is a Lie subalgebra of $\mathcal{V}(M)$: the Lie bracket $[V,W]$ of two vector fields V,W tangent to ∂M is again tangent to ∂M. Consequently, for $P \in \mathrm{Diff}_{\mathrm{b}}^{m}(M)$ and $Q \in \mathrm{Diff}_{\mathrm{b}}^{m'}(M)$:
(1) $P \circ Q \in \mathrm{Diff}_{\mathrm{b}}^{m+m'}(M)$;
(2) $[P,Q] \in \mathrm{Diff}_{\mathrm{b}}^{m+m'-1}(M)$.
If $\chi \in C^\infty(M)$, then $[P,\chi] \in \mathrm{Diff}_{\mathrm{b}}^{m-1}(M)$ with coefficients supported on $\operatorname{supp} \mathrm{d}\chi$.

The proof of Lemma 4.5, mostly bookkeeping, is found at the end of this section.

Using the local coordinate form Eq. (14), we can freeze the coordinates at $\rho = 0$ to obtain the *normal operator*

$$N(P) := \sum_{j=0}^{m} P_j(0)\left(\rho\partial_\rho\right)^j \in \mathrm{Diff}_{\mathrm{b}}^{m}([0,\varepsilon) \times \partial M).$$

Observe that $N(P)$ is dilation symmetric, like $r^2\Delta$ but unlike $a(r,y)r^2\Delta$. This is a strange situation: our manifold M generally has no global dilation symmetry, yet we have produced model operators that (formally) do. This was done by freezing the coefficients P_j. These P_j are smooth so that $P_j = P_j(0) + \mathcal{O}(\rho)$ and the difference between P and $N(P)$ vanishes at the boundary: near $\rho = 0$, every b-operator is, to leading order, indistinguishable from a dilation-invariant one. The geometric shape with a *global* dilation symmetry is a cone, an infinite cone to be precise. We now make a curious move. Having first attached a (non-physical) boundary representing infinity to our manifold, we now cut away everything but the boundary region and use that to define a model domain: the *model cone* $M_{\mathrm{b}} := [0,\infty)_\rho \times \partial M$. Near $\rho = 0$, M_{b} is indistinguishable from the collar of M; but M_{b} has a *global* dilation symmetry in ρ.

We now have a model to which the Mellin transform can be applied. In particular, the Mellin transform diagonalizes $N(P)$. This motivates the *indicial family* $N(P,\lambda) = \sum_{j=0}^{m} P_j(0)\lambda^j$ with, crucially,

$$\mathcal{M}[N(P)u](\lambda,\cdot) = N(P,\lambda)\tilde{u}(\lambda,\cdot).$$

Consider by analogy how freezing the coefficients at some interior point of a Riemannian manifold defines the Fourier symbol. Such a manifold is locally indistinguishable from a global translation-invariant $\mathbb{R}^n$. Fourier analysis can then be applied to study the Fourier symbol on $\mathbb{R}^n$ and local properties can be patched together into global properties via cutoff functions. Our approach will be essentially analogous. Recognizing the importance of the spectrum in functional analysis, we define the analogous *boundary spectrum* $\mathrm{spec}_{\mathrm{b}}(P) = \{\lambda \in \mathbb{C} \mid N(P,\lambda) \text{ is not invertible}\}$.

Definition 4.6. Let $P \in \mathrm{Diff}_{\mathrm{b}}^m(M)$, in the collar $P = \sum_{j=0}^m P_j(\rho)(\rho\partial_\rho)^j$.

- The *model cone* is $M_{\mathrm{b}} := [0,\infty)_\rho \times \partial M$.
- The *normal operator* is $N(P) := \sum_{j=0}^m P_j(0)(\rho\partial_\rho)^j$. It is an element of both $\mathrm{Diff}_{\mathrm{b}}^m(M)$ and $\mathrm{Diff}_{\mathrm{b}}^m(M_{\mathrm{b}})$.
- The *indicial family* is $N(P,\lambda) := \sum_{j=0}^m P_j(0)\lambda^j$.
- The *boundary spectrum* is $\mathrm{spec}_{\mathrm{b}}(P) := \{\lambda \in \mathbb{C} \mid N(P,\lambda) \text{ not invertible}\}$.

Applied to our standard example $r^2\Delta$, we see that $N(r^2\Delta,\lambda) = \lambda^2 - \lambda + \Delta_{\mathbb{S}^2}$. Noting that the eigenvalues of $\Delta_{\mathbb{S}^2}$ are $\{-\ell(\ell+1) \mid \ell \in \mathbb{N}_0\}$ where each eigenvalue corresponds to a spherical harmonic Y_ℓ, restricting to a particular harmonic gives $N(r^2\Delta,\lambda) = (\lambda-(\ell+1))(\lambda+\ell)$ so that $\mathrm{spec}_{\mathrm{b}}(r^2\Delta) = \{-\ell, \ell+1 \mid \ell \in \mathbb{N}_0\}$.

We also define a notion of b-ellipticity.

Definition 4.7 (b-Ellipticity)**.** A b-operator $P \in \mathrm{Diff}_{\mathrm{b}}^m(M)$ is *b-elliptic* if it is elliptic on the interior and in the collar

$$\sum_{j+|\alpha|=m} a_{j,\alpha}(\rho,y)\,\xi^j\eta^\alpha \neq 0 \tag{15}$$

for all $(\rho,y) \in M$ and all $(\xi,\eta) \in (\mathbb{R}\times\mathbb{R}^{n-1}) \setminus \{(0,0)\}$.

Eq. (15) gives a general definition of ellipticity. For a scalar operator of even order with real coefficients, this condition is equivalent to the operator's coefficient matrix being positive or negative definite, ensuring a bounded, coercive energy functional.

Lemma 4.6. Suppose $P \in \mathrm{Diff}_{\mathrm{b}}^m(M)$ is b-elliptic. Then:

(1) $N(P,\lambda) \in \mathrm{Diff}^m(\partial M)$ elliptic for all $\lambda \in \mathbb{C}$, its principal symbol is independent of λ.

(2) The formal adjoint $P^* \in \mathrm{Diff}_{\mathrm{b}}^m(M)$ and is also b-elliptic; its boundary spectrum is $-\overline{\mathrm{spec}_{\mathrm{b}}(P)}$.

Proof of Lemma 4.6. (1): In the collar, $N(P,\lambda) = \sum_{j=0}^m P_j(0)\lambda^j$ with $P_j(0) \in \mathrm{Diff}^{m-j}(\partial M)$. The order-$m$ part in ∂_y therefore comes from $j=0$, so the principal symbol of $N(P,\lambda)$ is

$$\sigma_m(N(P,\lambda))(y,\eta) = \sum_{|\alpha|=m} a_{0,\alpha}(0,y)\,(\mathrm{i}\eta)^\alpha,$$

independent of λ. Setting $\xi = 0$ in Eq. (15) shows it is nonvanishing for $\eta \neq 0$, i.e. $N(P,\lambda)$ is elliptic.

(2): We take adjoints with respect to the b-density $\mu = \rho^{-1}\,\mathrm{d}\rho\,\mathrm{dvol}_{h(0)}$. It suffices to compute adjoints of the generators of $\mathrm{Diff}_{\mathrm{b}}^m(M)$. Integration by parts gives, for u,v supported in the collar,

$$(\rho\partial_\rho)^* = -\rho\partial_\rho, \qquad V^* = -V + \mathrm{div}_{h(0)} V \quad (V \in \mathcal{V}(\partial M)), \qquad f^* = \bar{f},$$

where we used that μ is invariant under dilations in ρ. Since adjoints of products reverse order, $P^* \in \mathrm{Diff}_{\mathrm{b}}^m(M)$, with principal part $(-1)^m$ times the complex conjugate of that of P. In particular, Eq. (15) holds for P^* and P^* is b-elliptic.

For the boundary spectrum, note that freezing coefficients at $\rho = 0$ commutes with the computation above: the divergence corrections are smooth, so they freeze to their boundary values. Writing $P = \sum_j P_j(\rho)(\rho\partial_\rho)^j$ and taking adjoints termwise,

$$P^* = \sum_{j=0}^m (-\rho\partial_\rho)^j P_j(\rho)^*, \quad \text{hence} \quad N(P^*,\lambda) = \sum_{j=0}^m (-\lambda)^j P_j(0)^* = N(P,-\bar{\lambda})^*,$$

the last adjoint taken in $L^2(\partial M)$. An operator on ∂M is invertible if and only if its adjoint is, so $N(P^*,\lambda)$ is invertible iff $N(P,-\bar{\lambda})$ is, giving $\operatorname{spec}_{\mathrm{b}}(P^*) = -\overline{\operatorname{spec}_{\mathrm{b}}(P)}$. □

Proof of Lemma 4.5. Tangency is a local condition at the boundary, hence we can restrict our attention to the collar of M. Here, a b-vector field V is of the form $V_0 + V_1\rho\partial_\rho$ with $V_0 : [0,\varepsilon)_\rho \to \mathcal{V}(\partial M)$ a smooth family of vector fields on ∂M depending on ρ and $V_1 \in C^\infty([0,\varepsilon)_\rho \times \partial M)$. Consider W of the same form. We need to show that $[V,W]$ is also of this form i.e. that $[V,W] = U_0 + U_1\rho\partial_\rho$ for $U_0 : [0,\varepsilon)_\rho \to \mathcal{V}(\partial M)$ and $U_1 \in C^\infty([0,\varepsilon)_\rho \times \partial M)$. To this end, we expand $[V,W]$, giving

$$[V_0,W_0] + [V_0,W_1\rho\partial_\rho] + [V_1\rho\partial_\rho,W_0] + [V_1\rho\partial_\rho,W_1\rho\partial_\rho]$$

and consider each term separately.

- $[V_0,W_0]$. The Lie bracket of two smooth vector fields is a smooth vector field, hence this term contributes to U_0.
- $[V_0,W_1\rho\partial_\rho]$ and $[V_1\rho\partial_\rho,W_0]$. We apply $[A,fB] = (Af)B + f[A,B]$ to get

  $$[V_0,W_1\rho\partial_\rho] = V_0(W_1\rho)\,\partial_\rho + W_1\rho[V_0,\partial_\rho].$$

  Since V_0 is tangent to ∂M, $V_0(\rho) = 0$, so the first term is $V_0(W_1)\cdot\rho\partial_\rho$, contributing to U_1. Since ∂_ρ commutes with vector fields on ∂M, $[V_0,\partial_\rho] = -\partial_\rho V_0$, and the second term is $-W_1\rho(\partial_\rho V_0)$, which contributes to U_0. The same analysis applies to $[V_1\rho\partial_\rho,W_0]$.

  $V_0(\rho)$ does *not* mean 'evaluate V_0 at ρ'; it means 'V_0 applied to the coordinate function ρ.' Since V_0 is tangent to ∂M, this vanishes.
- $[V_1\rho\partial_\rho,W_1\rho\partial_\rho]$. We apply $[fA,gA] = f(Ag)A - g(Af)A$ to get

  $$[V_1\rho\partial_\rho,W_1\rho\partial_\rho] = (V_1\rho\partial_\rho W_1 - W_1\rho\partial_\rho V_1)\rho\partial_\rho.$$

  The coefficient is a smooth function of ρ hence it contributes to U_1.

Hence, all terms either contribute to U_0 or U_1 and $[V,W] \in \mathcal{V}_{\mathrm{b}}$.

Concerning (1), P and Q are by definition of the form

$$\sum_{j=1}^{m} a_j V_{j,1}V_{j,2}\cdots V_{j,k_j} \qquad k_j \le m.$$

$P\circ Q$ is a locally finite sum of products of $\le m+m'$ b-vector fields, hence in $\operatorname{Diff}_{\mathrm{b}}^{m+m'}(M)$.

Concerning (2), we expand

$$[P,Q] = \sum_{j,l}[a_jV_{j,1}\cdots V_{j,k_j},\, b_lW_{l,1}\cdots W_{l,k'_l}] \tag{16}$$

and apply $[fA,gB] = fg[A,B] + f(Ag)B - g(Bf)A$ to the terms of this sum: each equals $a_jb_l\,[V_{j,1}\cdots V_{j,k_j},\, W_{l,1}\cdots W_{l,k'_l}]$ plus correction terms of the form $a_j(Ab_l)B$ and $b_l(Ba_j)A$. Since Ab_l and Ba_j are smooth functions, the correction terms are smooth coefficients times a product of $k'_l \le m'$ or $k_j \le m$ b-vector fields respectively, so the corrections lie in $\operatorname{Diff}_{\mathrm{b}}^{m+m'-1}(M)$. It remains to show that $[V_{j,1}\cdots V_{j,k_j},\, W_{l,1}\cdots W_{l,k'_l}] \in \operatorname{Diff}_{\mathrm{b}}^{m+m'-1}(M)$. Applying $[AB,C] = A[B,C] + [A,C]B$ we can expand the main term into a sum of terms of the form $[V_i,W_j]$ multiplied by $m+m'-2$ b-vector fields on either side. Since $[V_i,W_j] \in \mathcal{V}_{\mathrm{b}}(M)$, each term is a product of $m+m'-1$ b-vector fields and the full bracket is a finite sum of such terms, it is thus in $\operatorname{Diff}_{\mathrm{b}}^{m+m'-1}(M)$.

For the last claim: $\chi \in C^\infty(M) \subset \mathrm{Diff}_\mathrm{b}^0(M)$, so (2) gives $[P,\chi] \in \mathrm{Diff}_\mathrm{b}^{m-1}(M)$. Concerning the support, consider Eq. (16) with $Q = \chi$. Each term becomes

$$[a_j V_{j,1} \cdots V_{j,k_j}, \chi] = a_j \sum_{i=1}^{k_j} V_{j,1} \cdots V_{j,i-1} (V_{j,i}\chi) V_{j,i+1} \cdots V_{j,k_j},$$

so every term carries a factor $V_{j,i}\chi$, which has support on $\operatorname{supp} \mathrm{d}\chi$. □

IV.2 b-Sobolev spaces

To do analysis on b-operators we need function spaces appropriate for analysis of boundary behavior. So far we did not refer to a metric: the b-operators are defined on a smooth manifold with boundary. To do analysis in L^2-spaces, however, we need a measure, and we need a metric to consider adjoints and ellipticity constants. Recall that we went from Δ to $r^2\Delta$ to obtain a dilation-invariant operator. The corresponding rescaling of the Euclidean metric is $\rho^2 g = (\mathrm{d}\rho/\rho)^2 + g_{\mathbb{S}^2}$. Generalizing, a *b-metric* on M is a metric which in the collar takes the form $g_\mathrm{b} = (\mathrm{d}\rho/\rho)^2 + h(\rho)$, with $h(\rho)$ a smooth family of metrics on ∂M. Its volume density $\mu = \rho^{-1}\,\mathrm{d}\rho\,\mathrm{d}y$ is invariant under dilations in ρ, and $L^2(M,\mu)$ is the pairing with respect to which the formal adjoint of a b-operator is again a b-operator (Lemma 4.6).

Definition 4.8. Let M be a compact manifold with boundary, ρ a boundary defining function and μ be a scale-invariant measure. Define

$$L_\mathrm{b}^{2,\alpha}(M) := \rho^\alpha L^2(M,\mu).$$

For $s \in \mathbb{N}_0$, the *b-Sobolev space* $H_\mathrm{b}^{s,\alpha}(M)$ is defined as

$$\left\{ u \in L_\mathrm{b}^{2,\alpha}(M) \;\middle|\; Pu \in L_\mathrm{b}^{2,\alpha}(M) \text{ for all } P \in \mathrm{Diff}_\mathrm{b}^s(M) \right\}. \tag{17}$$

An important corollary of Plancherel's identity in Fourier analysis is that $H^k(\mathbb{R})$ can be characterized via the norm in the Fourier domain: $H^k(\mathbb{R}) := \{ u \in L^2(\mathbb{R}) \mid \langle k\rangle^s \hat{u} \in L^2(\mathbb{R})\}$. The following lemma is analogous.

Lemma 4.7. $H_\mathrm{b}^{s,\alpha}(\mathbb{R}_+) = \{ u \in L^{2,\alpha}(\mathbb{R}_+, \frac{\mathrm{d}x}{x}) \mid \langle\tau\rangle^s \mathcal{M}[u](\alpha + \mathrm{i}\tau) \in L^2(\mathbb{R}) \}$, i.e. the norms

$$\|u\|^2_{H_\mathrm{b}^{s,\alpha}(\mathbb{R}_+)} \quad \text{and} \quad \int_\mathbb{R} \langle\tau\rangle^{2s} \left|\mathcal{M}[u](\alpha+\mathrm{i}\tau)\right|^2 \mathrm{d}\tau$$

are equivalent. Consequently, $H_\mathrm{b}^{s,\alpha}(\mathbb{R}_+)$ is defined for $s \in \mathbb{R}$.

This lemma follows by the change of coordinates we have been using throughout the chapter. However, since we will be working with general manifolds, it will be insightful to state the effect of such a coordinate change more formally. The (compactification of the) domain $\mathbb{R}_+$ is then the special case of a manifold with degenerate ∂M i.e. where ∂M is a single point.

Lemma 4.8. The coordinate change $\rho \mapsto e^{-t}$ sends a collar $[0,\varepsilon)_\rho \times \partial M$ to a half-cylinder $[T,\infty)_t \times \partial M$ with $T = -\log\varepsilon$. Under this change of variables:

$$\rho\partial_\rho \longleftrightarrow -\partial_t, \qquad \rho^\alpha \longleftrightarrow e^{-\alpha t}, \qquad \rho^{-1}\,\mathrm{d}\rho\,\mathrm{d}y \longleftrightarrow \mathrm{d}t\,\mathrm{d}y.$$

Consequently, $P \in \mathrm{Diff}_{\mathrm{b}}^m(M)$ becomes a standard differential operator of order m on the cylinder; multiplication by ρ^α becomes multiplication by $e^{-\alpha t}$; and $L_{\mathrm{b}}^{2,\alpha}(M)$ restricted to the collar becomes $e^{-\alpha t}L^2([T,\infty)\times\partial M)$. In particular, $H_{\mathrm{b}}^{s,\alpha}$ on the collar corresponds to $e^{-\alpha t}H^s([T,\infty)\times\partial M)$.

This lemma allows us to establish properties of b-Sobolev spaces via their standard Sobolev counterpart.

Proposition 4.7. The following properties transfer from standard Sobolev theory. Let $P \in \mathrm{Diff}_{\mathrm{b}}^m(M)$. Let $\chi_{\mathrm{b}}, \chi_{\mathrm{int}} \in C^\infty(M)$.

$\chi_{\mathrm{b}}, \chi_{\mathrm{int}} \in C^\infty(M)$ are such that $\chi_{\mathrm{b}} + \chi_{\mathrm{int}} = 1$; χ_{b} supported and $\equiv 1$ in a collar of ∂M. Then $\operatorname{supp} \mathrm{d}\chi_{\mathrm{b}} = \operatorname{supp} \mathrm{d}\chi_{\mathrm{int}}$ is compact in M°.

(1) In the interior, $H_{\mathrm{b}}^{s,\alpha} = H_{\mathrm{loc}}^s$; the b-spaces differ from standard Sobolev only at the boundary, where they allow controlled growth/decay specified by the weight α.
(2) $u \in H_{\mathrm{b}}^{s,\alpha}(M)$ if and only if $\chi_{\mathrm{int}}u \in H^s(M^\circ)$ and the Mellin transform $\widetilde{\chi_{\mathrm{b}}u}(\alpha+\mathrm{i}\tau, y)$ satisfies

$$\int_{\mathbb{R}} \|(1+\tau^2-\Delta_{\partial M})^{s/2}\widetilde{\chi_{\mathrm{b}}u}(\alpha+\mathrm{i}\tau,\cdot)\|_{L^2(\partial M)}^2 \,\mathrm{d}\tau < \infty.$$

These together define a norm equivalent to $\|u\|_{H_{\mathrm{b}}^{s,\alpha}}$. Consequently, $H_{\mathrm{b}}^{s,\alpha}(M)$ is defined for $s \in \mathbb{R}$.
(3) $P : H_{\mathrm{b}}^{s,\alpha}(M) \to H_{\mathrm{b}}^{s-m,\alpha}(M)$ is a bounded operator. In particular, multiplication of $u \in H_{\mathrm{b}}^{s,\alpha}$ by functions in $C^\infty(M)$ is bounded.
(4) For $\chi \in \{\chi_{\mathrm{int}}, \chi_{\mathrm{b}}\}$, the commutator $[P,\chi] : H_{\mathrm{b}}^{s,\alpha}(M) \to H_{\mathrm{b}}^{s-m+1,\alpha-N}(M)$ is bounded for every $N \in \mathbb{R}$.
(5) The inclusion $H_{\mathrm{b}}^{s,\alpha}(M) \hookrightarrow H_{\mathrm{b}}^{s',\alpha'}(M)$ is continuous when $s \geq s'$ and $\alpha \geq \alpha'$, and compact when $s > s'$ *and* $\alpha > \alpha'$.
(6) Pairing via the L_{b}^2 inner product identifies $H_{\mathrm{b}}^{s,\alpha}$ and $H_{\mathrm{b}}^{-s,-\alpha}$ as duals.

Proof. Items (4) and (5) deserve attention, the others are almost immediate in light of Lemma 4.8: (1) reduces to the fact that on compact $K \subset M^\circ$, the weight ρ^α and the b-derivatives are equivalent to their standard counterparts up to constants depending on K; (2) follows from (1) together with Lemma 4.7 applied on the collar; (3) follows from the boundedness of standard differential operators with bounded coefficients on $H^s(\mathbb{R}\times\partial M)$; lastly, (6) reduces to the duality $(H^s)^* = H^{-s}$.

(4): $[P,\chi] \in \mathrm{Diff}_{\mathrm{b}}^{m-1}(M)$ by Lemma 4.5. That $[P,\chi] : H_{\mathrm{b}}^{s,\alpha} \to H_{\mathrm{b}}^{s-m+1,\alpha}$ is bounded follows from (3). For $N \leq 0$ the claim then follows from the continuous inclusion (5). Suppose $N > 0$. The coefficients of $[P,\chi]$ are supported in $\operatorname{supp}\mathrm{d}\chi$ (Lemma 4.5). By the support property, the support lies in $\{\rho \geq \rho_0\}$ for some $\rho_0 > 0$, hence $\operatorname{supp}[P,\chi]u \subset \{\rho \geq \rho_0\}$. In this region $\rho^{-(\alpha-N)} \leq \rho^{-\alpha}\rho_0^N$, giving $\|[P,\chi]u\|_{H_{\mathrm{b}}^{s-m+1,\alpha-N}} \leq \rho_0^N\|[P,\chi]u\|_{H_{\mathrm{b}}^{s-m+1,\alpha}} \leq C_N\|u\|_{H_{\mathrm{b}}^{s,\alpha}}$, using the case $N = 0$ just proven.

(5): The continuous inclusion is obvious. Note that a mere change of variables is not sufficient for compactness, since $H^s \hookrightarrow H^{s'}$ is *not* compact on the non-compact half-cylinder $[T,\infty)\times\partial M$. However, $H^s \hookrightarrow H^{s'}$ is compact on the restricted domain $[T,T_0] \subset [T,\infty)$. We show that this suffices by controlling the tail $[T_0,\infty)$. Throughout, we identify functions on M with their cylindrical counterparts in $H^{s'}$ and the interval $[T,T_0]$ with the cylindrical region $[T,T_0]\times\partial M$.

Let $(u_n) \subset H_{\mathrm{b}}^{s,\alpha}(M)$ be a bounded sequence, and let $T_k \to \infty$. We identify a nested family of subsequences as follows. On the compact region $[T, T_1] \times \partial M$, there exists a subsequence $(u_n^{(1)}) \subset (u_n)$ converging in $H^{s'}$. Iteratively, on $[T, T_{k+1}] \times \partial M$, identify $(u_n^{(k+1)}) \subset (u_n^{(k)})$ converging in $H^{s'}$. The diagonal sequence $v_k := u_k^{(k)}$ then converges in $H^{s'}$ on every compact region $[T, T_N] \times \partial M$: for $k \geq N$, v_k lies in $(u_n^{(N)})$, which is Cauchy on $[T, T_N]$. It remains to show that (v_k) is Cauchy in $H_{\mathrm{b}}^{s',\alpha'}(M)$. Let χ_k be a cutoff supported on $[T, T_k] \times \partial M$. Split

$$v_k - v_\ell = \chi_N(v_k - v_\ell) + (1 - \chi_N)(v_k - v_\ell),$$

for $N \leq k, \ell$ to be chosen. On $\operatorname{supp}(1 - \chi_N) \subset [T_N, \infty) \times \partial M$, the weight ratio gives

$$\|(1-\chi_N)w\|_{H_{\mathrm{b}}^{s',\alpha'}} \leq e^{-(\alpha-\alpha')T_N} \|(1-\chi_N)w\|_{H_{\mathrm{b}}^{s',\alpha}} \lesssim e^{-(\alpha-\alpha')T_N} \|w\|_{H_{\mathrm{b}}^{s,\alpha}}$$

for any w. Applied to $v_k - v_\ell$, and using boundedness of (v_k) in $H_{\mathrm{b}}^{s,\alpha}$,

$$\|(1-\chi_N)(v_k - v_\ell)\|_{H_{\mathrm{b}}^{s',\alpha'}} \leq C e^{-(\alpha-\alpha')T_N} < \varepsilon$$

for sufficiently large N. For $k, \ell \geq N$, both v_k and v_ℓ lie in the N-th subsequence, which converges in $H^{s'}$ on the compact region $\operatorname{supp} \chi_N$. Hence, $\|\chi_N(v_k - v_\ell)\|_{H_{\mathrm{b}}^{s',\alpha'}} < \varepsilon$ for k, ℓ sufficiently large. Thus (v_k) is Cauchy in $H_{\mathrm{b}}^{s',\alpha'}(M)$, and the embedding is compact. □

IV.3 Asymptotic analysis

Difficult problems often stem from imprecise statements; once formulated clearly, the solution is often straightforward. What we have done so far provides the framework for stating our problem precisely. We have developed the Mellin transform, defined the domain, the operators, and the function spaces. With these in hand, the problem we wish to solve is to determine the asymptotic behavior of $P : H_{\mathrm{b}}^{s,\alpha}(M) \to H_{\mathrm{b}}^{s',\alpha'}(M)$ where M is a compact manifold with boundary and a P is a b-operator. However, what does it mean to 'determine asymptotic behaviour'? To have a precise statement of the problem, we need to clarify what counts as a satisfactory solution.

The motivating observation at the start of the chapter was that differential operators have characteristic asymptotic modes, and that we can express a solution by decomposing the data into such modes. The (weighted) Laplacian has integral asymptotic modes, each corresponding to a spherical harmonic into which the boundary data is decomposed. Hence, if we have identified the modes of an operator—its boundary spectrum—then a satisfactory answer could be the identification of how 'present' each mode is our data. The Mellin transform of the normal operator gives us these modes as poles of a meromorphic function, their residue measures the presence of the mode. Consequently, if there are *finitely* many poles present, then Mellin identifies them and we are done. However, in general, the indicial family of a b-operator has infinitely many roots (accumulating at $\operatorname{Re} z = \pm\infty$). We thus proceed sequentially in order of increasing $\operatorname{Re} z_j$, i.e. from slowest to fastest decay. One might hope for a pointwise equality $u(x) = \sum_{j=1}^{\infty} c_j x^{z_j} (\log x)^{k_j}$ in the limit. However, this is false: the coefficients c_j may grow fast enough that the sum diverges. For example, in Stirling's approximation

$$n! \sim \sqrt{2\pi n}\,\left(\tfrac{n}{e}\right)^n \exp\left(\sum_{k=1}^{\infty} \frac{B_{2k}}{2k(2k-1)\,n^{2k-1}}\right), \tag{18}$$

Eq. (18) arises from shifting the contour in a Mellin–Barnes representation of $\Gamma(z)$; for the optimal truncation see Paris and Kaminski 2001, §6.4.

the infinite series diverges for every fixed n, because the Bernoulli numbers B_{2k} grow factorially. Yet, truncating the series at any N gives an asymptotic equivalence as $n \to \infty$, with the optimal $N \approx \pi n$ giving a discrepancy exponentially small in n. The factorial $n!$ is perfectly well-defined for every n, and the divergent series describes its asymptotic behavior exactly, just not as an equality. This is not a defect of Stirling's formula but a feature of the question: we are not asking what $n!$ *equals*, we are asking what $n!$ is *asymptotically*. These are categorically different questions. The relation between a function and its asymptotic expansion is a controlled discrepancy at each finite order, not equality.

What does 'controlled discrepancy' mean in our case? Suppose we collect modes up to some weight γ_2 and call everything beyond it the remainder. If the remainder is of strictly faster decay, if it is dominated by every mode we kept, then we know its size *exactly*! This is just the contour-integration argument of Theorem 4.2: suppose the inverse Mellin transform of $\mathcal{M}[u]$ along $\operatorname{Re} s = \gamma_1$ recovers u; deforming to $\operatorname{Re} s = \gamma_2$ picks up the residues at the poles in between, so that u equals these residues plus the contour integral along $\operatorname{Re} s = \gamma_2$. The residues are the asymptotic modes and the other contour is the remainder, which is of strictly faster decay. This deserves a more formal statement.

Definition 4.9. Let $\chi_{\mathrm{b}} \in C_c^{\infty}([0,\infty))$ be a fixed cutoff. For a finite index set $\mathcal{E} \subset \mathbb{C} \times \mathbb{N}_0$, define the finite-dimensional space of asymptotic modes

$$\mathcal{P}_{\mathcal{E}} := \operatorname{span}\{\,\chi_{\mathrm{b}}(\rho)\,\rho^z(\log\rho)^k \mid (z,k) \in \mathcal{E}\}.$$

Lemma 4.9. Let $\gamma_1 < \gamma_2$ and let $\mathcal{E}$ be a finite index set supported in $S_{(\gamma_1,\gamma_2)}$. If $u \in H_{\mathrm{b}}^{s,\gamma_1}(\mathbb{R}_+)$ is such that $\mathcal{M}[u]$ extends meromorphically to the strip $S_{[\gamma_1,\gamma_2]}$ with poles described by $\mathcal{E}$ and $\langle\tau\rangle^s \mathcal{M}[u](\gamma_2 + \mathrm{i}\tau) \in L^2(\mathbb{R})_{\tau}$, then

$$u \in \mathcal{P}_{\mathcal{E}} \oplus H_{\mathrm{b}}^{s,\gamma_2}(\mathbb{R}_+).$$

The proof is immediate from Corollary 4.2.2.

We can apply the lemma iteratively: each application extends the strip rightward, identifies new modes, and shrinks the remainder. If this can be done for arbitrarily large N—if u decomposes into modes plus a remainder in $H_{\mathrm{b}}^{s,N}$—then we call u *polyhomogeneous of regularity s*.

Definition 4.10. Let $\mathcal{E} \subset \mathbb{C} \times \mathbb{N}_0$ be an index set. A function u is *polyhomogeneous of regularity s*, written $u \in \mathcal{A}_{\mathrm{phg}}^{s;\mathcal{E}}(M)$, if for every $N \in \mathbb{R}$,

$$u \in \mathcal{P}_{\mathcal{E}_N} \oplus H_{\mathrm{b}}^{s,N}(M);$$

$\mathcal{E}_N := \{(z,k) \in \mathcal{E} : \operatorname{Re} z < N\}$

i.e. if there exist $c_{z,k} \in \mathbb{C}$ such that

$$u - \sum_{(z,k)\in\mathcal{E}_N} c_{z,k}\,\chi_{\mathrm{b}}(\rho)\,\rho^z(\log\rho)^k \in H_{\mathrm{b}}^{s,N}(M).$$

The particular kind of index set, determined by the b-elliptic operator, that we will need in our main theorem is as follows.

Definition 4.11. Given a b-elliptic P, let $\mathcal{E}(P) \subset \mathbb{C} \times \mathbb{N}_0$ be the smallest set containing $\{(z,k) \mid z \text{ a pole of } N(P,\lambda)^{-1} \text{ of order } > k\}$ and closed under the following two operations:
(1) if $(z,k) \in \mathcal{E}(P)$, then $(z+1,k) \in \mathcal{E}(P)$;
(2) if $(z,k) \in \mathcal{E}(P)$ and z is a pole of $N(P,\lambda)^{-1}$ of order p, then $(z,k+j) \in \mathcal{E}(P)$ for $j < p$.

Our method is now set: Mellin identifies the modes, contour integration measures the discrepancy at each cutoff, and iterating indefinitely determines a polyhomogeneous expansion.

We pause to indicate how the theory so far relates to the literature.

(1) Hintz 2023 considers, in essence, the case that everything is smooth. If the data and coefficients are smooth, then elliptic-regularity arguments give solutions in $H_{\mathrm{b}}^{\infty,\alpha} := \bigcap_{s\in\mathbb{N}} H_{\mathrm{b}}^{s,\alpha}$. These functions are essentially what is called *conormal*: smooth functions with $\mathcal{O}(\rho^\alpha)$ decay at the boundary under all $\mathcal{V}_{\mathrm{b}}$-derivatives. The two notions differ only by an ε-loss weight, the precise statement being as follows.

Lemma 4.10. Define the conormal functions on M of weight α as

$$\mathcal{A}^\alpha(M) := \{u \in C^\infty(M^\circ) \mid \forall P \in \bigcup_{m\in\mathbb{N}_0} \mathrm{Diff}_{\mathrm{b}}^m(M),\ \exists C : |Pu| \le C\rho^\alpha\}.$$

We have

$$H_{\mathrm{b}}^{\infty,\alpha}(M) \subset \mathcal{A}^\alpha(M) \subset \bigcap_{\varepsilon>0} H_{\mathrm{b}}^{\infty,\alpha-\varepsilon}(M).$$

Cf. Hintz 2023, Proposition 2.14

Proof. First inclusion: let $u \in H_{\mathrm{b}}^{\infty,\alpha}(M)$ and $P \in \mathrm{Diff}_{\mathrm{b}}^m(M)$, so that $Pu \in H_{\mathrm{b}}^{s,\alpha}(M)$ for every s (Proposition 4.7(3)). For $s > \dim(M)/2$, the Sobolev embedding $H^s \subset C^0$ applies on the cylinder $[T,\infty)_t \times \partial M$ with a uniform constant, so Lemma 4.8 gives $\sup|\rho^{-\alpha}Pu| \le C\|Pu\|_{H_{\mathrm{b}}^{s,\alpha}}$, i.e. $|Pu| \le C\rho^\alpha$. Hence $u \in \mathcal{A}^\alpha(M)$. Second inclusion: let $u \in \mathcal{A}^\alpha(M)$, $P \in \mathrm{Diff}_{\mathrm{b}}^m(M)$, $\varepsilon > 0$. Then $|\rho^{-(\alpha-\varepsilon)}Pu| \le C\rho^\varepsilon$, and $\int_0^1 \rho^{2\varepsilon}\,\frac{\mathrm{d}\rho}{\rho} < \infty$, so $Pu \in L_{\mathrm{b}}^{2,\alpha-\varepsilon}(M)$. Since P was arbitrary, $u \in H_{\mathrm{b}}^{\infty,\alpha-\varepsilon}(M)$. □

The ε cannot be removed: $\rho^\alpha \in \mathcal{A}^\alpha(M)$, but $1 \notin L^2(\frac{\mathrm{d}\rho}{\rho})$ near $\rho = 0$, so $\rho^\alpha \notin H_{\mathrm{b}}^{\infty,\alpha}(M)$.

Polyhomogeneous functions are then defined exactly as in Definition 4.10, with the remainder in $\mathcal{A}^N$ instead of $H_{\mathrm{b}}^{s,N}$. These spaces agree, by the lemma, with our $\mathcal{A}_{\mathrm{phg}}^{\infty,\mathcal{E}}(M)$ modulo the ε weight.

(2) Schulze 1991 takes a different approach. Our problems are defined by two pieces of data: u lives a priori in some b-Sobolev space and the operator defines a set of asymptotic modes $\mathrm{spec}_{\mathrm{b}}(P)$. Schulze packages these two into a single function space. We first give a simplified form of his definition, adapted by us, that connects to our theory so far: let $H_{\mathrm{b},\mathcal{E}}^{s,\gamma}(\mathbb{R}_+)$ denote the functions in the b-Sobolev space $H_{\mathrm{b}}^{s,\gamma}$ whose Mellin transform extends meromorphically to a strip with poles described by some set $\mathcal{E}$. The decomposition of Lemma 4.9 becomes the structural property of this space, which he expresses by means of a short exact sequence:

$$0 \longrightarrow H_{\mathrm{b}}^{s,\gamma}(\mathbb{R}_+) \longrightarrow H_{\mathrm{b},\mathcal{E}}^{s,\gamma}(\mathbb{R}_+) \longrightarrow \mathcal{Z}_{\mathcal{E}} \longrightarrow 0,$$

Exactness means that the first map is injective, the second is surjective, and its kernel is the image of the first. Thus the quotient $H_{\mathrm{b},\mathcal{E}}^{s,\gamma}/H_{\mathrm{b}}^{s,\gamma}$ is identified with $\mathcal{Z}_{\mathcal{E}}$.

here $\mathcal{Z}_P$ is the space of Laurent coefficients of the poles in $\mathcal{E}$.

Schulze's actual definition uses *open* strips with uniform L^2 control on every interior vertical line in the Mellin domain rather than L^2 on a particular boundary line. This avoids committing to whether the boundary line is pole-free and serves a larger goal: constructing an operator algebra that is closed under composition, supports parametrix construction, and extends to manifolds with edges and higher singularities. The open-strip formulation with uniform L^2 control on every interior vertical line affords flexibility under these operations, where asymptotic data can shift or accumulate. Even multiplication by a smooth function $a(\rho, y)$ generates infinitely many shifted poles (via Taylor expansion in ρ), forcing a shadow condition on the index sets $\mathcal{E}$. As a price, his spaces are Fréchet rather than Banach. Since our work is restricted to a single Fredholm theorem at chosen weights, we do not need this flexibility and can work with L^2-norms on particular boundary lines.

The shadow condition closes $\mathcal{E}$ under the integer shifts generated by Taylor powers ρ^j, see Schulze 1991, p. 49 and compare Definition 4.11.

Definition 4.12. Let $H^{k,\gamma}_{\mathrm{b},P}(\mathbb{R}_+)_\Delta$ be the space of those $u \in L^{2,\gamma}(\mathbb{R}_+)$ with the following two properties:

(1) For all $\varepsilon > 0$ for all $\beta \in S^\gamma_{\Delta_\varepsilon}$

$$\langle s\rangle^k \chi(s)\mathcal{M}[u](s) \in L^2(\operatorname{Re} s = \beta)$$

where $\chi(s)$ is a $\pi_{\mathbb{C}}P$-excision function. More strongly, we require a uniform bound $\sup_\beta \|\chi u\|_{k,\beta} < \infty$ (norm defined below).

(2) $\mathcal{M}[u](s)$ is meromorphic in a strip S^γ_Δ and $(z_j, k_j) \in P \cap S^\gamma_\Delta$ describe its poles with location z_j and multiplicity k_j.

The associated seminorms are twofold: corresponding to (1) are

$$\|\chi u\|_{k,\beta} := \|\langle s\rangle^k \chi(s)\mathcal{M}[u](s)\|_{L^2(\operatorname{Re} s=\beta)}$$

with variable β and χ; and corresponding to (2) are $h \mapsto |\zeta_{j,k}|$, where $\zeta_{j,k}$ is the coefficient in the Laurent series of h at $(z - p_j)^{-(k+1)}$, $p_j \in S^\sigma_\Delta$.

$\Delta := (-\delta', \delta)$ for some $\delta', \delta \in \mathbb{R}$ and $\Delta_\varepsilon := (\delta + \varepsilon, \delta' - \varepsilon)$; $S^\gamma_\Delta := \{z \in \mathbb{C} \mid \operatorname{Re} z \in \Delta - \gamma\}$; a A-excision function is a function $\chi \in C^\infty(\mathbb{C})$ taking values in $[0,1]$ so that: $\chi(z) = 0$ for all z in an open neighbourhood of A and $\chi(z) = 1$ outside another open neighbourhood of A.

A second virtue of this definition is that it generalizes to what Schulze calls *continuous asymptotics*, where the discrete index set P is replaced by a closed subset $V \subset \mathbb{C}$, the Mellin transform of u having pole-like singularities in V rather than merely discrete non-accumulating poles. The asymptotic expansion of u is correspondingly written as an integral over V rather than a discrete sum. A motivating application is edges: on a manifold with an edge Y, the Mellin poles of the indicial family depend on the edge variable $y \in Y$, tracing out trajectories in $\mathbb{C}$ that may bifurcate or intersect. A discrete index set cannot parametrize this, but a closed set V can absorb continuous trajectories.

This generalization is considered in Schulze 1991, §1.2.2.

A manifold with edges is locally a smooth family of cones: the cone tips trace out the edge Y, while away from Y the space is an ordinary smooth manifold, see Schulze 1991, p.266.

(3) Lesch 1997 considers the same object: his *Mellin symbol* $\sigma^{\mu,\nu}_M(P)(z) = \sum_k A_k(0)z^k$ (his Def. 1.1.4) is our indicial family $N(P,\lambda)$, and his $\operatorname{spec}\sigma^{\mu,\nu}_M(P)$ is our boundary spectrum, with the same finite-order poles of the inverse (Prop. & Def. 1.1.9). He in fact proves a more general Fredholm theorem than ours, for *all* closed extensions of an elliptic Fuchs-type operator (Prop. 1.3.16). However, rather than pursuing the polyhomogeneous expansion of solutions (as in Theorem 4.3), he extracts spectral invariants—heat-trace asymptotics (Thm. 2.4.6), ζ- and η-functions

(Ch. V), and an index theorem (Thm. 2.4.8)—through the heat kernel and the Brüning–Seeley singular asymptotics lemma (§2.1).

V Elliptic operators on compactified manifolds

With the theory in place and a clearly identified problem, we are able to state and prove the theorem that generalizes the gravitational example we started with.

Theorem 4.3. Let $P \in \mathrm{Diff}_{\mathrm{b}}^{m}(M)$ be an elliptic *b*-operator on a compact manifold M with boundary ∂M. Let $N(P,\lambda)$ be its indicial family. Then, for $s > m$ and $\alpha \notin \mathrm{Re}\left(\mathrm{spec}_{\mathrm{b}}(P)\right)$:

- $P : H_{\mathrm{b}}^{s,\alpha}(M) \to H_{\mathrm{b}}^{s-m,\alpha}(M)$ is Fredholm.
- If $Pu = f \in \dot{\mathcal{C}}^{\infty}(M)$ then u is polyhomogeneous with asymptotic modes determined by the poles of $N(P,\lambda)^{-1}$.

$\dot{\mathcal{C}}^{\infty}(M)$ are the smooth function that are flat at ∂M, i.e. the function and all its derivatives are zero there.

Lemma 4.11. Under the same conditions as Theorem 4.3:

(1) The set $\mathrm{spec}_{\mathrm{b}}(P) \subset \mathbb{C}$ is discrete, locally finite, and coincides with the set of poles of the meromorphic family $\lambda \mapsto N(P,\lambda)^{-1}$, each pole of finite order.

(2) $N(P,\lambda) : H^{m}(\partial M) \to L^{2}(\partial M)$ is invertible for all $\lambda \in \mathbb{C} \setminus \mathrm{spec}_{\mathrm{b}}(P)$.

(3) For any $\alpha \in \mathbb{R}$ such that the vertical line $\mathrm{Re}\,\lambda = \alpha$ avoids $\mathrm{spec}_{\mathrm{b}}(P)$, for all $\lambda = \alpha + i\tau$:

$$\left(\sum_{j=0}^{m} \langle \tau \rangle^{2(m-j)} \|\tilde{v}\|_{H^{j}(\partial M)}^{2} \right)^{1/2} \leq C_{\alpha} \|N(P,\alpha+i\tau)\tilde{v}\|_{L^{2}(\partial M)}$$

Let us first sketch how Theorem 4.3 will be proven. We would like a semi-Fredholm estimate

$$\|u\|_{X} \leq \|Pu\|_{Y} + \|Ru\|_{Z} \tag{19}$$

with $R : X \to Z$ compact. This implies that P has a finite dimensional kernel and closed range (Proposition A.1) and Fredholmness follows by extending the same argument to the adjoint. The main obstacle is that our candidate for $R : X \to Z$ is $H_{\mathrm{b}}^{s,\alpha} \hookrightarrow H_{\mathrm{b}}^{s',\alpha'}$, which requires *both* $s > s'$ and $\alpha > \alpha'$ for compactness. We proceed with a divide-and-conquer strategy. First, use classical elliptic regularity to justify $s > s'$; second, distinctive to the b-calculus, we use the model operator $N(P)$ to justify $\alpha > \alpha'$, while controlling $P - N(P)$. Together they give the semi-Fredholm estimate. Within each step, we partition u into the interior and the collar, requiring us to account for commutators $[P,\chi]$. The polyhomogeneous expansion is subsequently constructed using Lemma 4.11.

Proof of Lemma 4.11. We consider $N(P,\lambda)$ as a holomorphic family of differential operators on ∂M. Since each $N(P,\lambda)$ is elliptic (Lemma 4.6) and ∂M is compact, the family consists of Fredholm operators $H^{m}(\partial M) \to L^{2}(\partial M)$ (Theorem A.1). If we can construct a single λ_0 at which $N(P,\lambda_0)$ is invertible, then by the analytic Fredholm theorem (Theorem A.3), $\lambda \mapsto N(P,\lambda)^{-1}$is meromorphic on $\mathbb{C}$ with poles of finite order at a discrete set which is $\mathrm{spec}_{\mathrm{b}}(P)$. Discreteness gives finiteness on compact sets and finiteness in vertical strips follows once we note that the estimate established below holds with constants uniform for $\sigma \in [a,b]$, so that $N(P,\sigma+\mathrm{i}\tau)$ is invertible for $|\tau| \geq T_0(a,b)$, proving (1).

Part (2) holds by definition of $\mathrm{spec}_\mathrm{b}(P)$. Given (1) and (2), the inequality of (3) restricted to $\{|\tau| \leq T_0\}$ follows from continuity of $\tau \mapsto N(P,\sigma+\mathrm{i}\tau)^{-1}$ on this compact set, since $\mathrm{Re}\,\lambda = \sigma$ avoids $\mathrm{spec}_\mathrm{b}(P)$ by hypothesis and $\langle\tau\rangle$ is bounded there.

It remains to establish the estimate (3) for $|\tau| > T_0$ and to construct λ_0, which we accomplish simultaneously. To this end, consider, in local coordinates,

$$N(P,\lambda) = \sum_{j+|k|\leq m} a_{j,k}(0,y)\lambda^j \partial_y^k.$$

The joint principal symbol of $N(P,\lambda)$ in (η,λ) is

$$p(y;\eta,\lambda) = \sum_{j+|k|=m} a_{j,k}(0,y)\lambda^j (\mathrm{i}\eta)^k.$$

First, observe that p is homogeneous of degree m, i.e. $p(y;t\eta,t\lambda) = t^m p(y;\eta,\lambda)$. On the imaginary axis $\lambda = \mathrm{i}\tau$ we have $p(y;\eta,\mathrm{i}\tau) = \mathrm{i}^m \sum_{j+|k|=m} a_{j,k}\tau^j\eta^k$, which is i^m times the b-symbol of the ellipticity condition (Eq. (15)) at real (τ,η). b-ellipticity thus gives $p(y;\eta,\mathrm{i}\tau) \neq 0$ for $(\tau,\eta) \in \mathbb{R}^n \setminus 0$. By homogeneity and compactness of $\{\tau^2 + |\eta|^2 = 1\} \times \partial M$,

$$|p(y;\eta,\lambda)| \geq c(|\eta|^2 + |\tau|^2)^{m/2} \quad \text{uniformly in } y \in \partial M.$$

(1) *Frozen coefficients.* Fix $y_0 \in \partial M$ and let $N_0(\lambda)$ denote $N(P,\lambda)$ with coefficients frozen at y_0. Its Fourier symbol on $\mathbb{R}^{n-1}$ is of the form

$$\hat{N}_0(\eta,\lambda) = p(y_0;\eta,\lambda) + r(y_0;\eta,\lambda),$$

where r denotes all terms with $j+|k| < m$. Since $|r|^2 \leq C(|\eta|^2+|\tau|^2)^{(m-1)}$, we can choose an $R \in \mathbb{R}$ large enough that $|p|$ dominates $|r|$, giving

$$|\hat{N}_0(\eta,\lambda)| \geq \frac{c}{2}(|\eta|^2+|\tau|^2)^{m/2} \quad \text{for } |\eta|^2+|\tau|^2 \geq R^2 \tag{20}$$

and by Plancherel,

$$\|N_0(\lambda)u\|^2_{L^2(\mathbb{R}^{n-1})} = \int_{\mathbb{R}^{n-1}} |\hat{N}_0(\eta,\lambda)|^2 |\hat{u}(\eta)|^2 \,\mathrm{d}\eta.$$

Taking $|\tau| \geq \max(R,1)$ so that $\tau^2 \geq \frac{1}{2}\langle\tau\rangle^2$ (which we will need later) and $\tau^2 + |\eta|^2 \geq R^2$ for every η, so that Eq. (20) applies globally:

$$\|N_0(\lambda)u\|^2_{L^2} \geq \frac{c^2}{4} \int_{\mathbb{R}^{n-1}} (|\eta|^2+|\tau|^2)^m |\hat{u}(\eta)|^2 \,\mathrm{d}\eta.$$

Since

$$(|\eta|^2+|\tau|^2)^m \geq 2^{-m} \sum_{j=0}^m \binom{m}{j} \langle\tau\rangle^{2(m-j)} |\eta|^{2j},$$

applying Plancherel termwise gives

$$\sum_{j=0}^m \langle\tau\rangle^{2(m-j)} \|u\|^2_{H^j} \leq C\|N_0(\sigma+\mathrm{i}\tau)u\|^2_{L^2}. \tag{21}$$

This is the large-τ estimate with frozen coefficients.

(2) *Variable coefficients, locally.* On a chart $U \subset \partial M$ with coordinates in $\mathbb{R}^{n-1}$, locally decompose $N(P,\lambda) = N_0(\lambda) + R(y,\lambda)$. By taking U small enough, the coefficients of R are bounded by ε uniformly and $|\lambda|^2 = \sigma^2 + \tau^2 \leq C_{[a,b]}\langle\tau\rangle^2$ for $\sigma \in [a,b]$. Hence, for u supported in U,

$$\|R(\lambda)u\|^2_{L^2(\mathbb{R}^{n-1})} \leq \varepsilon \sum_{j=0}^{m} \langle\tau\rangle^{2(m-j)} \|u\|^2_{H^j(\mathbb{R}^{n-1})}.$$

Then from the frozen coefficients bound Eq. (21) we get

$$\|N(P,\lambda)u\|^2_{L^2} \geq \frac{1}{2}\|N_0(\lambda)u\|^2_{L^2} - \|R(\lambda)u\|^2_{L^2} \geq \left(\frac{1}{2C} - \varepsilon\right) \sum_{j=0}^{m} \langle\tau\rangle^{2(m-j)} \|u\|^2_{H^j}.$$

Taking $\varepsilon < 1/2C$ gives an inequality of the same form as Eq. (21), with $N(P,\lambda)$ instead of $N_0(\lambda)$.

(3) *Variable coefficients, globally.* Cover ∂M by finitely many small charts $\{U_\ell\}$ and consider the subordinate partition of unity $\{\chi_\ell\}$. For each ℓ, apply the local estimate of item (2) to $\chi_\ell \tilde{v}$:

$$\sum_{j=0}^{m} \langle\tau\rangle^{2(m-j)} \|\chi_\ell \tilde{v}\|^2_{H^j} \leq C_1 \|N(P,\lambda)(\chi_\ell \tilde{v})\|^2_{L^2}.$$

Write

$$N(P,\lambda)(\chi_\ell \tilde{v}) = \chi_\ell N(P,\lambda)\tilde{v} + [N(P,\lambda),\chi_\ell]\tilde{v}.$$

Now, $[N(P,\lambda),\chi_\ell]$ has order $\leq m-1$ in ∂_y and degree $\leq m-1$ in λ (Lemma 4.5 applied pointwise in λ):

$$\|[N(P,\lambda),\chi_\ell]\tilde{v}\|^2_{L^2} \leq C \sum_{j=0}^{m-1} \langle\tau\rangle^{2(m-j-1)} \|\tilde{v}\|^2_{H^j}.$$

The commutator contribution $\sum_j \langle\tau\rangle^{2(m-j-1)}\|\tilde{v}\|^2_{H^j}$ has each term dominated by $\langle\tau\rangle^{-2}$ times its counterpart $\langle\tau\rangle^{2(m-j)}\|\tilde{v}\|^2_{H^j}$. For $|\tau| \geq T_0$ sufficiently large, the commutator contribution can therefore be subtracted from its counterpart, giving the global estimate

$$\sum_{j=0}^{m} \langle\tau\rangle^{2(m-j)} \|\tilde{v}\|^2_{H^j(\partial M)} \leq C \|N(P,\sigma + \mathrm{i}\tau)\tilde{v}\|^2_{L^2(\partial M)}.$$

This is the estimate for large τ; in particular, $N(P,\sigma+\mathrm{i}\tau)$ is injective on $H^m(\partial M)$ for $|\tau| \geq T_0$. The formal adjoint $N(P,\lambda)^* = N(P^*,-\bar{\lambda})$ has the same principal symbol (Lemma 4.6), so the same argument gives T_0' with $N(P,\sigma+\mathrm{i}\tau)^*$ injective on $H^m(\partial M)$ for $|\tau| \geq T_0'$. Setting $\lambda_0 := \sigma + i\max(T_0, T_0')$, if $v \in L^2(\partial M)$ annihilates $\operatorname{ran} N(P,\lambda_0)$, then $N(P,\lambda_0)^* v = 0$ in the sense of distributions. Elliptic regularity on the closed manifold ∂M gives $v \in C^\infty(\partial M) \subset H^m(\partial M)$, hence $v = 0$ by the injectivity of $N(P,\lambda_0)^*$. The range of $N(P,\lambda_0)$ is closed (Fredholmness) with trivial annihilator, hence the range is $L^2(\partial M)$ so that λ_0 is the sought point of invertibility. □

Proof of Theorem 4.3. We will use cutoffs $\chi_{\mathrm{b}}, \chi_{\mathrm{int}} \in C^\infty(M)$.

$\chi_{\mathrm{b}}, \chi_{\mathrm{int}} \in C^\infty(M)$ are such that $\chi_{\mathrm{b}} + \chi_{\mathrm{int}} = 1$; χ_{b} supported and $\equiv 1$ in a collar of ∂M. Then $\operatorname{supp} \mathrm{d}\chi_{\mathrm{b}} = \operatorname{supp} \mathrm{d}\chi_{\mathrm{int}}$ is compact in M°.

(1) *Elliptic regularity.* We analyse the interior and the collar separately, using $\|u\|_{H^{s,\alpha}_{\mathrm{b}}(M)} \leq \|\chi_{\mathrm{b}} u\|_{H^{s,\alpha}_{\mathrm{b}}(M)} + \|\chi_{\mathrm{int}} u\|_{H^{s,\alpha}_{\mathrm{b}}(M)}$.

Interior: $\chi_{\text{int}} u \in H^s_{\text{loc}}(M^\circ)$ (Proposition 4.7 (1)) and b-ellipticity on the interior coincides with standard ellipticity. Hence, we apply standard interior regularity (Theorem A.1) to get

$$\|\chi_{\text{int}} u\|_{H^{s,\alpha}_{\text{b}}} \le C(\|P\chi_{\text{int}} u\|_{H^{s-m,\alpha}_{\text{b}}} + \|\chi_{\text{int}} u\|_{H^{0,\alpha}_{\text{b}}}). \tag{22}$$

Collar: we use the local coordinates and the standard coordinate change $\rho = e^{-t}$ of Lemma 4.8 to convert our problem to a classical Sobolev problem involving t. Moreover, under this change of coordinates, the uniform b-elliptic operator becomes a uniformly elliptic operator. This time we use the uniformity to get an elliptic estimate (Theorem A.2), in the original coordinates,

$$\|\chi_{\text{b}} u\|_{H^{s,\alpha}_{\text{b}}} \le C(\|P\chi_{\text{b}} u\|_{H^{s-m,\alpha}_{\text{b}}} + \|\chi_{\text{b}} u\|_{H^{0,\alpha}_{\text{b}}}). \tag{23}$$

To combine Eqs. (22) and (23) into a single estimate on u, we expand $P\chi_i u = \chi_i P u + [P, \chi_i]u$ and bound their $H^{s-m,\alpha}_{\text{b}}$-norms: the first by continuity of multiplication by a smooth function (Proposition 4.7 (3)), and the second by applying Proposition 4.7 (4) with $s-1$ in place of s, giving

$$\|[P,\chi_i]u\|_{H^{s-m,\alpha}_{\text{b}}} \le C\|u\|_{H^{s-1,\alpha}_{\text{b}}}.$$

Using the triangle inequality we conclude

$$\|u\|_{H^{s,\alpha}_{\text{b}}(M)} \le C\left(\|Pu\|_{H^{s-m,\alpha}_{\text{b}}(M)} + \|u\|_{H^{s-1,\alpha}_{\text{b}}(M)}\right). \tag{24}$$

(2) *Inversion of the indicial family.* We first consider the model $N(P)$ on the model cone $(\rho, y) \in M_{\text{b}}$. Our goal is to bound functions $\chi_{\text{b}} u$ supported in the collar neighborhood which is diffeomorphic to a neighborhood of $\rho = 0$ in the model cone M_{b}; b-Sobolev norms agree under the identification of these two regions.

The Mellin transform of $\chi_{\text{b}} u$ with respect to ρ, fixed on the line $\operatorname{Re}\lambda = \alpha$, gives,

$$\widetilde{\chi_{\text{b}} u}(\tau, y) = \mathcal{M}[\chi_{\text{b}} u](\alpha + \mathrm{i}\tau, y).$$

Assuming $\alpha \notin \operatorname{Re}(\operatorname{spec}_{\text{b}}(P))$, then by Lemma 4.11

$$\sum_{j=0}^{m} \langle\tau\rangle^{2(m-j)} \|\widetilde{\chi_{\text{b}} u}(\tau,\cdot)\|^2_{H^j(\partial M)} \le C_\alpha \|N(P, \alpha + i\tau)\widetilde{\chi_{\text{b}} u}(\tau,\cdot)\|^2_{L^2(\partial M)}.$$

Multiplying by $\langle\tau\rangle^{2(s-m)}$ and integrating with respect to τ gives

$$\begin{aligned}\int_{\mathbb{R}} \sum_{j=0}^{m} \langle\tau\rangle^{2(s-j)} \|\widetilde{\chi_{\text{b}} u}(\tau,\cdot)\|^2_{H^j(\partial M)}\, \mathrm{d}\tau \\ \le C_\alpha \int_{\mathbb{R}} \langle\tau\rangle^{2(s-m)} \|N(P, \alpha + i\tau)\widetilde{\chi_{\text{b}} u}(\tau,\cdot)\|^2_{L^2(\partial M)}\, \mathrm{d}\tau.\end{aligned}$$

Since the Mellin transform diagonalizes the indicial operator, i.e. $N(P, \alpha + \mathrm{i}\tau)\widetilde{\chi_{\text{b}} u}(\tau,\cdot) = \mathcal{M}[N(P)(\chi_{\text{b}} u)](\alpha + \mathrm{i}\tau, \cdot)$, we apply Plancherel for Mellin (Proposition 4.7 (2)),

$$\|\chi_{\text{b}} u\|^2_{H^{s,\alpha}_{\text{b}}(M)} \le C_\alpha \|N(P)(\chi_{\text{b}} u)\|^2_{H^{s-m,\alpha}_{\text{b}}(M)}. \tag{25}$$

Turning to the actual operator P, we seek to apply our model estimate Eq. (25) on $N(P)$ while controlling the discrepancy. Starting from the identity

$$\chi_{\mathrm{b}}Pu = N(P)(\chi_{\mathrm{b}}u) + (P - N(P))(\chi_{\mathrm{b}}u) - [P, \chi_{\mathrm{b}}]u,$$

we rearrange, fill it into the RHS of Eq. (25) and use the triangle inequality to get

$$\|\chi_{\mathrm{b}}u\|_{H_{\mathrm{b}}^{s,\alpha}} \leq C\Big(\|\chi_{\mathrm{b}}Pu\|_{H_{\mathrm{b}}^{s-m,\alpha}} + \underbrace{\|(P-N(P))(\chi_{\mathrm{b}}u)\|_{H_{\mathrm{b}}^{s-m,\alpha}}}_{(a)} + \underbrace{\|[P,\chi_{\mathrm{b}}]u\|_{H_{\mathrm{b}}^{s-m,\alpha}}}_{(b)}\Big).$$

We analyse the two discrepancies (a) and (b) separately.

(a): $P - N(P) \in \rho\mathrm{Diff}_{\mathrm{b}}^m$ by Taylor expansion of the coefficients $P_j(\rho)$. The operator $P - N(P) : H_{\mathrm{b}}^{s,\alpha-1} \to H_{\mathrm{b}}^{s-m,\alpha}$ is bounded (Proposition 4.7(3)). Hence, (a) is bounded by $\|\chi_{\mathrm{b}}u\|_{H_{\mathrm{b}}^{s,\alpha-1}} \leq C\|u\|_{H_{\mathrm{b}}^{s,\alpha-1}}$.

(b): $[P,\chi_{\mathrm{b}}]$ is in $\mathrm{Diff}_{\mathrm{b}}^{m-1}$ (Lemma 4.5) and is bounded (Proposition 4.7 (4)) so that (b) is bounded by $C_N\|u\|_{H_{\mathrm{b}}^{s-1,\alpha-N}}$ for any N. Absorbing the bound on (b) into that on (a) gives

$$\|\chi_{\mathrm{b}}u\|_{H_{\mathrm{b}}^{s,\alpha}} \leq C\left(\|Pu\|_{H_{\mathrm{b}}^{s-m,\alpha}} + \|u\|_{H_{\mathrm{b}}^{s,\alpha-1}}\right). \tag{26}$$

We also need a bound on $\chi_{\mathrm{int}}u$. On interior compact sets, b-Sobolev norms with different weights are equivalent (Proposition 4.7 (1)). Hence,

$$\|u\|_{H_{\mathrm{b}}^{s,\alpha}} \leq \|\chi_{\mathrm{b}}u\|_{H_{\mathrm{b}}^{s,\alpha}} + C_N\|\chi_{\mathrm{int}}u\|_{H_{\mathrm{b}}^{s,-N}} \quad \text{for any } N.$$

Combining gives a global estimate

$$\|u\|_{H_{\mathrm{b}}^{s,\alpha}} \leq C\left(\|Pu\|_{H_{\mathrm{b}}^{s-m,\alpha}} + \|u\|_{H_{\mathrm{b}}^{s,\alpha-1}}\right). \tag{27}$$

Applying Eq. (24) to the remainder of Eq. (27) gives the semi-Fredholm estimate

$$\|u\|_{H_{\mathrm{b}}^{s,\alpha}(M)} \leq C\Big(\|Pu\|_{H_{\mathrm{b}}^{s-m,\alpha}(M)} + \|u\|_{H_{\mathrm{b}}^{s-1,\alpha-1}(M)}\Big). \tag{28}$$

Hence, P has a finite-dimensional kernel and closed range (Proposition A.1).

(3) *Fredholm property.* To conclude that P is Fredholm, we have to show that the cokernel is finite dimensional. Since $\operatorname{ran}(P)$ is closed, the closed range theorem gives $\operatorname{coker}(P) \cong \operatorname{ran}(P)^{\perp}$. By Proposition 4.7(6) this annihilator is in $H_{\mathrm{b}}^{-(s-m),-\alpha}$. An element v of this dual annihilates $\operatorname{ran}(P)$ if and only if $\langle Pu, v\rangle_{L_{\mathrm{b}}^2} = 0$ for all u, equivalently $\langle u, P^*v\rangle_{L_{\mathrm{b}}^2} = 0$ for all u, i.e. $P^*v = 0$. Hence

$$\operatorname{coker}(P) \cong \ker(P^* : H_{\mathrm{b}}^{-(s-m),-\alpha} \to H_{\mathrm{b}}^{-s,-\alpha}).$$

By Lemma 4.6, $P^* \in \mathrm{Diff}_{\mathrm{b}}^m(M)$, P^* is uniformly b-elliptic and $\operatorname{spec}_{\mathrm{b}}(P^*) = -\overline{\operatorname{spec}_{\mathrm{b}}(P)}$ so that $\alpha \notin \operatorname{Re}\left(\operatorname{spec}_{\mathrm{b}}(P^*)\right)$. Hence, the analysis of steps (1) and (2) applies to P^*. We thus obtain a similar estimate

$$\|u\|_{H_{\mathrm{b}}^{-(s-m),-\alpha}} \leq C\left(\|P^*u\|_{H_{\mathrm{b}}^{-s,-\alpha}} + \|u\|_{H_{\mathrm{b}}^{-(s-1),-(\alpha-1)}}\right),$$

and $\ker(P^*) \cong \operatorname{coker}(P)$ is finite dimensional. We conclude that P is Fredholm.

(4) *Polyhomogeneous expansion.* $Pu = f \in \dot{\mathcal{C}}^\infty(M) \subset H_{\mathrm{b}}^{s-m,\alpha}(M)$ for every s. Hence, $u \in H_{\mathrm{b}}^{s,\alpha}(M)$ for every s, i.e. $u \in H_{\mathrm{b}}^{\infty,\alpha}(M)$. Now consider $\chi_{\mathrm{b}} u$ as a function on the model cone M_{b}. We start with

$$P(\chi_{\mathrm{b}} u) = \chi_{\mathrm{b}} f + [P, \chi_{\mathrm{b}}]u.$$

Since the commutator $[P, \chi_{\mathrm{b}}]$ is supported on the interior, so $[P, \chi_{\mathrm{b}}]u \in \dot{\mathcal{C}}^\infty$. Hence,

$$N(P)(\chi_{\mathrm{b}} u) = \chi_{\mathrm{b}} f + [P, \chi_{\mathrm{b}}]u + (N(P) - P)(\chi_{\mathrm{b}} u) =: g,$$

Using the diagonalization $\mathcal{M}[N(P)v] = N(P,\lambda)\tilde{v}$,

$$N(P,\lambda)\,\widetilde{\chi_{\mathrm{b}} u}(\lambda, \cdot) = \tilde{g}(\lambda, \cdot).$$

Since $\alpha \notin \operatorname{Re}\big(\operatorname{spec}_{\mathrm{b}}(P)\big)$, Lemma 4.11(2) gives that $N(P,\lambda)$ is invertible on $\operatorname{Re}\lambda = \alpha$, so

$$\widetilde{\chi_{\mathrm{b}} u}(\lambda, \cdot) = N(P,\lambda)^{-1}\tilde{g}(\lambda, \cdot) \quad \text{on } \operatorname{Re}\lambda = \alpha.$$

By Lemma 4.11(1), $N(P,\lambda)^{-1}$ is meromorphic with poles of finite order at $\operatorname{spec}_{\mathrm{b}}(P)$. The term $(N(P) - P)(\chi_{\mathrm{b}} u)$ lies in $\rho \cdot H_{\mathrm{b}}^{\infty,\alpha} = H_{\mathrm{b}}^{\infty,\alpha+1}$ since $P - N(P) \in \rho\mathrm{Diff}_{\mathrm{b}}^m(M)$; its Mellin transform is therefore holomorphic on $\operatorname{Re}\lambda < \alpha + 1$ with vertical Schwartz decay on vertical lines there (Propositions 4.2 and 4.4). The other contributions to g are in $\dot{\mathcal{C}}^\infty(M)$, so their Mellin transforms are entire with vertical Schwartz decay. Hence $\widetilde{\chi_{\mathrm{b}} u} = N(P,\lambda)^{-1}\tilde{g}$ extends meromorphically to $\operatorname{Re}\lambda < \alpha + 1$, with poles at $\operatorname{spec}_{\mathrm{b}}(P)$ of order bounded by that of $N(P,\lambda)^{-1}$, i.e. poles described by $\mathcal{E}(P)$.

We iterate with u replaced by the remainder. By Lemma 4.9 applied on the strip up to some pole-free line $\gamma \in (\alpha, \alpha+1)$, $\chi_{\mathrm{b}} u = \varphi + u_1$ with $\varphi \in \mathcal{P}_{\mathcal{E}(P)_\gamma}$ and $u_1 \in H_{\mathrm{b}}^{\infty,\gamma}$. In the identity defining g, the error term now splits as $(N(P) - P)\varphi + (N(P) - P)u_1$. The second lies in $H_{\mathrm{b}}^{\infty,\gamma+1}$ as before. For the first, $P - N(P)$ has smooth coefficients, so Taylor expansion in ρ shows that applied to a mode $\rho^z(\log\rho)^k$ it gives an asymptotic sum of modes $\rho^{z+j}(\log\rho)^{k'}$ with $j \geq 1$, $k' \leq k$, modulo $H_{\mathrm{b}}^{\infty,N}$ for any N: shifted modes, with poles as prescribed by operation (1) of Definition 4.11. Applying $N(P,\lambda)^{-1}$ can increase the order of a pole only where a shifted pole lies in $\operatorname{spec}_{\mathrm{b}}(P)$, by the order of $N(P,\lambda)^{-1}$ there: this is operation (2) of Definition 4.11. Hence at every stage the poles of $\widetilde{\chi_{\mathrm{b}} u}$ are described by $\mathcal{E}(P)$. Each iteration advances the strip of meromorphy by one unit, so finitely many iterations extend past any fixed N, with vertical Schwartz decay away from the poles by the polynomial growth of $N(P,\lambda)^{-1}$ on pole-free lines (Lemma 4.11(3)).

Fix $N \in \mathbb{R}$. By discreteness of $\operatorname{spec}_{\mathrm{b}}(P)$, pick $\gamma_N > N$ such that the line $\operatorname{Re}\lambda = \gamma_N$ contains no point of $\mathcal{E}(P)$. On the strip S_{α,γ_N}, the function $\widetilde{\chi_{\mathrm{b}} u}$ is meromorphic with discrete, locally finite poles, and its holomorphic part has vertical Schwartz decay. Applying Corollary 4.2.3 gives

$$\chi_{\mathrm{b}} u(\rho, y) = \sum_{(z,k) \in \mathcal{E}_{\gamma_N}} c_{z,k}(y)\,\chi_{\mathrm{b}}(\rho)\,\rho^z(\log\rho)^k + r_N(\rho, y), \qquad r_N \in H_{\mathrm{b}}^{\infty,\gamma_N}.$$

Since $\chi_{\mathrm{int}} u$ is supported in the interior, where u is smooth, it contributes nothing to the asymptotic behavior at ∂M. Letting N increase indefinitely culminates in the polyhomogeneous expansion

$$u(\rho,y) \sim \sum_{(z,k)\in\mathcal{E}(P)} c_{z,k}(y)\chi(\rho)\rho^z(\log\rho)^k.$$

$\square$

We thus see that the general theme of this thesis encompasses this very different setting as well: the 'simple modes' in which the solution is to be decomposed—a polyhomogeneous expansion with the modes determined by the boundary spectrum—are not merely determined by the differential operator but also by the nature of the geometric domain. Concretely, we can hold the operator constant, say the Laplacian, and vary the shape. Consider a metric cone $g = \mathrm{d}r^2 + r^2 g_Y$ over a closed manifold (Y, g_Y) of dimension $n-1$, compactified at infinity with boundary defining function $\rho = 1/r$ as before. In the collar,

$$r^2\Delta = (\rho\partial_\rho)^2 - (n-2)\,\rho\partial_\rho + \Delta_Y$$

is b-elliptic with boundary spectrum $\left\{\frac{1}{2}\left((n-2) \pm \sqrt{(n-2)^2 + 4\mu}\right) \,\middle|\, \mu \in \operatorname{spec}(-\Delta_Y)\right\}$. Hence, three domains:

- the compactification $\overline{\mathbb{R}^3}$, where $Y = \mathbb{S}^2$ gives $\mu = l(l+1)$ and $\operatorname{spec}_\mathrm{b} = \{l+1, -l\}$: the integers, the decay rates Laplace found in section I;
- the global monopole, asymptotically the cone over a sphere of radius a, where $\mu = l(l+1)/a^2$ and $\operatorname{spec}_\mathrm{b} = \left\{\frac{1}{2}\left(1 \pm \sqrt{1+4l(l+1)/a^2}\right)\right\}$: one shape parameter, and the spectrum slides off the integers;
- the cylindrical end $g = \mathrm{d}t^2 + g_Y$, where $\rho = e^{-t}$ gives $\rho\partial_\rho = -\partial_t$ (Lemma 4.8) and $\operatorname{spec}_\mathrm{b} = \{\pm\sqrt{\mu_k}\}$: the first-order term has vanished, and the spectrum is symmetric about zero.

The three domains correspond to three spectra: integers, surds, square roots. These spectra determine how solutions are to be decomposed and consequently what is observable in the asymptotic regime.

The technique of compactifying the noncompact $\mathbb{R}^3$ for the sake of asymptotic analysis as $|x| \to \infty$ also extends to the analysis of a different kind of singularity. The operator

$$P := \Delta + \frac{1}{|x|} = r^{-2}\left((r\partial_r)^2 + r\partial_r + \Delta_{\mathbb{S}^2} + r\right)$$

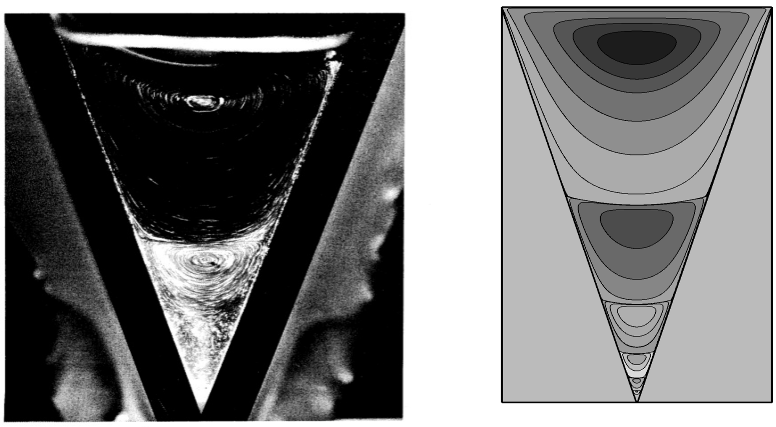
is a (weighted) b-operator and it can be considered on the *blow-up* of $\mathbb{R}^3$ at $\{0\}$, i.e. $[\mathbb{R}^3; \{0\}] := [0,\infty) \times \mathbb{S}^2$. The same techniques we have developed for the compactification of $\mathbb{R}^3$ then apply to the boundary at $r = 0$.

See Hintz 2023, Ch.5.

Conclusion

We will conclude this thesis, not with a summary, but with an example. This example considers the abstract theory of Chapter 4 in a more concrete setting, thereby drawing out the general theme of 'geometrical shape as a cause'. Since our goal here is to stress the conceptual meaning of the mathematics, we will be more informal with the computations in this section compared to Chapter 4.

When a viscous fluid flows past a sharp corner, a sequence of progressively smaller vortices (called eddies) forms at the corner, each successive one spinning in opposite direction. The flow pattern does not change in time, but the fluid inside the eddies is constantly circulating. Each eddy is progressively smaller, so that you only observe the first few. This phenomenon is known as Moffatt eddies (see Fig. 2).



(a) (b)

FIGURE 2. Moffatt eddies in a triangular cavity: (a) experimental flow visualization and (b) numerical stream-function contours at $Re = 1$. The eddies alternate in direction and decrease in size toward the acute corner. Reproduced from Biswas and Kalita 2017, Fig. 3.

We define a wedge region

$$W_{\alpha,R} = \{(r,\theta) \mid r \in (0,R), -\alpha < \theta < \alpha\}. \tag{29}$$

We denote the velocity with $u = (u^r, u^\theta)$, the pressure p, and the viscosity $\mu > 0$. We consider the BVP steady state incompressible Stokes flow with

Dirichlet boundary conditions:

For a more systematic derivation of a model for the corner analysis see Branicki and Moffatt 2006.

$$\begin{cases} -\nabla p + \mu \Delta u = 0, \quad \nabla \cdot u = 0 & \text{in } W_{\alpha,R}, \\ u(r, \pm\alpha) = 0 & 0 < r < R, \\ u(R,\theta) = g(\theta) & -\alpha < \theta < \alpha. \end{cases} \tag{30}$$

We first reformulate this using a so-called stream function $\psi : W_{\alpha,R} \to \mathbb{R}$. The stream function is an analogue of the potential function, defined such that

For incompressible rather than irrotational flows.

$$u^r = \frac{1}{r}\partial_\theta \psi, \qquad u^\theta = -\partial_r \psi. \tag{31}$$

It transforms the problem into the following BVP

$$\begin{cases} \Delta^2 \psi = f & \text{in } W_{\alpha,\infty}, \\ \partial_\theta \psi(r,\pm\alpha) = 0, \quad \partial_r \psi(r,\pm\alpha) = 0 & 0 < r < \infty. \end{cases} \tag{32}$$

Here we have absorbed the BC at $r = R$ into a function $f(r,\theta)$. This means we are working on a dilation-symmetric model cone so that this problem can be analysed with our Mellin techniques.

Aware of the dilation symmetry with respect to r, we seek to study Δ^2 through a related b-operator, the *weighted* biharmonic equation $r^4\Delta^2$. The first step is to describe it as a polynomial of $r\partial_r$ and ∂_θ. However, since r^2 and Δ do not commute, we cannot simply take the square of the expression of the weighted Laplacian $r^2\Delta$. We use a conjugation trick to obtain

Hintz 2023, Lemma 2.3

$$r^4\Delta^2 = \left((r\partial_r - 2)^2 + \partial_\theta^2\right)\left((r\partial_r)^2 + \partial_\theta^2\right). \tag{33}$$

Taking the Mellin transform of this operator gives

$$N(s)\tilde{v}(s,\theta) := \left((s-2)^2 + \partial_\theta^2\right)\left(s^2 + \partial_\theta^2\right)\tilde{v}(s,\theta).$$

The radial BC $\partial_r \psi(r,\pm\alpha) = 0$ transforms after integration by parts into $(s+1)\tilde{\psi}(s+1,\pm\alpha) = 0$ for all s, which is equivalent to $\tilde{\psi}(s,\pm\alpha) = 0$. Hence, we have transformed the BVP Eq. (32) into

$$\begin{cases} N(s)\tilde{v}(s,\theta) = \tilde{f}(s) \\ \partial_\theta \tilde{v}(s,\pm\alpha) = 0 \\ \tilde{v}(s,\pm\alpha) = 0. \end{cases} \tag{34}$$

For each fixed s, Eq. (34) is a constant-coefficient ODE in θ. By standard ODE theory, its homogeneous solutions are exponentials $e^{\mu\theta}$ with μ a root of the characteristic polynomial of $N(s)$. The four roots are $\mu \in \{\pm \mathrm{i}s,\ \pm \mathrm{i}(s-2)\}$, hence the basis

$$\phi_1 = e^{\mathrm{i}s\theta}, \qquad \phi_2 = e^{-\mathrm{i}s\theta}, \qquad \phi_3 = e^{\mathrm{i}(s-2)\theta}, \qquad \phi_4 = e^{-\mathrm{i}(s-2)\theta}$$

forms the solution space of $N(s)\tilde{v} = 0$. The general solution is

$$\tilde{v}(s,\theta) = \sum_{k=1}^{4} c_k(s)\,\phi_k(\theta),$$

with the coefficients $c_k(s)$ determined by the BC at $\theta = \pm\alpha$. Imposing the BC gives a 4×4 linear system for the c_k. The values of s at which this system fails to be invertible define $\mathrm{spec}_{\mathrm{b}}(r^4\Delta^2)$. It therefore consists of the zeros of the determinant of the 4×4 matrix $M(s)$ obtained by evaluating the basis $\phi_1,\dots,\phi_4$ and their derivatives at $\theta=\pm\alpha$:

$$\det\begin{pmatrix} e^{\mathrm{i}s\alpha} & e^{-\mathrm{i}s\alpha} & e^{\mathrm{i}(s-2)\alpha} & e^{-\mathrm{i}(s-2)\alpha} \\ e^{-\mathrm{i}s\alpha} & e^{\mathrm{i}s\alpha} & e^{-\mathrm{i}(s-2)\alpha} & e^{\mathrm{i}(s-2)\alpha} \\ \mathrm{i}se^{\mathrm{i}s\alpha} & -\mathrm{i}se^{-\mathrm{i}s\alpha} & \mathrm{i}(s-2)e^{\mathrm{i}(s-2)\alpha} & -\mathrm{i}(s-2)e^{-\mathrm{i}(s-2)\alpha} \\ \mathrm{i}se^{-\mathrm{i}s\alpha} & -\mathrm{i}se^{\mathrm{i}s\alpha} & \mathrm{i}(s-2)e^{-\mathrm{i}(s-2)\alpha} & -\mathrm{i}(s-2)e^{\mathrm{i}(s-2)\alpha} \end{pmatrix}=0.$$

Analyzing a 4×4 determinant is rather tedious, but a shortcut is offered by considering the symmetries of the domain. The wedge $W_{\alpha,R}$ is invariant under the reflection $\theta\mapsto-\theta$ and both the operator and the boundary conditions inherit this symmetry. We therefore expect that we can decompose our solution into independent odd and even parts. Recombining the exponential basis accordingly gives

$$\cos(s\theta),\quad \cos((s-2)\theta),\quad \sin(s\theta),\quad \sin((s-2)\theta).$$

We can similarly translate the BC into independent conditions on the even and odd parts, so that the 4×4 system decouples into two independent 2×2 blocks

$$\det M(s)=\det M_{\text{even}}(s)\cdot\det M_{\text{odd}}(s).$$

In particular, the even block is

$$M_{\text{even}}(s)=\begin{pmatrix}\cos(s\alpha) & \cos((s-2)\alpha)\\ -s\sin(s\alpha) & -(s-2)\sin((s-2)\alpha)\end{pmatrix},$$

whose determinant is

$$-s\cos((s-2)\alpha)\sin(s\alpha)+(s-2)\cos(s\alpha)\sin((s-2)\alpha)=0,$$

which simplifies into

$$\sin(2\alpha(s-1))=-(s-1)\sin(2\alpha).$$

Applying a change of variables $x=2\alpha(s-1)$ leads to

$$\frac{\sin(x)}{x}=-\frac{\sin(2\alpha)}{2\alpha}. \tag{35}$$

A similar computation for the odd block gives

$$\frac{\sin x}{x}=\frac{\sin(2\alpha)}{2\alpha}.$$

These transcendental equations in x define the boundary spectrum of our BVP; they form $\operatorname{spec}_{\mathrm{b}}(r^4\Delta^2)$.

Because f is supported away from the corner, its Mellin transform is entire with rapid vertical decay (Proposition 4.4), so $\tilde\psi=N(s)^{-1}\tilde f$ is meromorphic on all of $\mathbb{C}$, with poles precisely at $\operatorname{spec}_{\mathrm{b}}(r^4\Delta^2)$. A single contour shift in the inversion integral, in the manner of Corollary 4.2.3, then converts each pole crossed into a term of the expansion: as $r\to 0$,

$$\psi(r,\theta)\sim\sum_{s_j\in\operatorname{spec}_{\mathrm{b}}} c_j r^{s_j} f_j(\theta),$$

where f_j is the angular profile produced by the residue of $N(s)^{-1}\tilde f$ at s_j, and simple poles are assumed for simplicity of notation. The expansion sums over only those s_j to the right of the weight line on which we invert, and in this case this line is determined by the physics: finite viscous dissipation near the corner requires $\nabla u\in L^2$, which for a mode

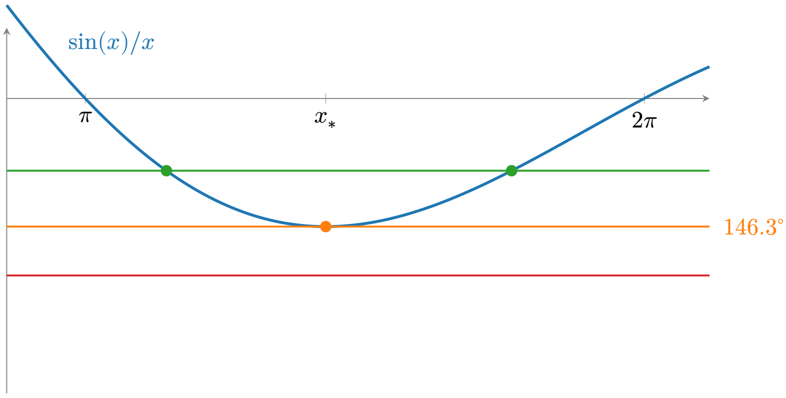


FIGURE 3. Graphical construction of the roots of Eq. (35). As the opening angle 2α decreases, the two real intersections coincide at $2\alpha \approx 146.3°$ and subsequently form a complex-conjugate pair.

r^s amounts to $\operatorname{Re} s > 1$. The asymptotics of ψ at the corner are therefore governed by the elements of the boundary spectrum with $\operatorname{Re} s_j > 1$, ordered by increasing real part.

As $r \to 0$ the behavior of ψ is governed by the eigenvalue with smallest real part larger than 1, say s_1, since the corresponding $r^{s_1} f_{s_1}(\theta)$ dominates the asymptotic expansion. If s_1 is real, then r^{s_1} is monotone in r; if s_1 is complex, then the factor $r^{\mathrm{i}\operatorname{Im} s_1} = e^{\mathrm{i}\operatorname{Im}(s_1)\ln r}$ oscillates infinitely often as $r \to 0$, producing a self-similar sequence of eddies—the Moffatt eddies. A graphical construction of Eq. (35) reveals that the leading eigenvalue is complex when

$$2\alpha \lessapprox 146.3°.$$

There is thus a threshold below which the shape of the corner leads to oscillatory behavior of ψ. In other words, the sharpness of a corner *causes* the Moffatt eddies.

Appendix A

Functional analysis

Theorem A.1. Let Y be a compact manifold without boundary and $P \in \mathrm{Diff}^m(Y)$ elliptic. Then $P : H^s(Y) \to H^{s-m}(Y)$ is Fredholm for every $s \in \mathbb{R}$, and for every $s_0 \in \mathbb{R}$ there is C with

$$\|u\|_{H^s} \le C(\|Pu\|_{H^{s-m}} + \|u\|_{H^{s_0}}).$$

See Hörmander 2007 Theorem 19.2.1

Theorem A.2. Let Y be a compact Riemannian manifold without boundary, and let $P \in \mathrm{Diff}^m(\mathbb{R} \times Y)$ be C^∞-bounded and uniformly elliptic. Then, for every $s, s_0 \in \mathbb{R}$, there exists $C > 0$ such that

$$\|u\|_{H^s(\mathbb{R}\times Y)} \le C\left(\|Pu\|_{H^{s-m}(\mathbb{R}\times Y)} + \|u\|_{H^{s_0}(\mathbb{R}\times Y)}\right)$$

for all $u \in H^s(\mathbb{R} \times Y)$.

See Shubin 1992, Lemma 1.4.

Definition A.1 (Fredholm operator)**.** A *Fredholm operator* is a bounded operator between Banach spaces $T : X \to Y$ with

(1) $\dim \ker T < \infty$,
(2) $\mathrm{ran} T \subset Y$ is closed,
(3) $\dim \mathrm{coker} T := \dim(Y/\mathrm{ran} T) < \infty$.

Proposition A.1. Let X, Y, Z be Banach spaces, $P \in \mathcal{L}(X, Y)$, and $R : X \to Z$ compact. Suppose there exists C such that

$$\|u\|_X \le C(\|Pu\|_Y + \|Ru\|_Z) \qquad \forall u \in X.$$

Then $\dim \ker P < \infty$ and $\mathrm{ran} P \subset Y$ is closed.

See Hintz 2025 Theorem 5.74.

Theorem A.3 (Analytic Fredholm Theorem)**.** Let X, Y be Banach spaces, let $\Omega \subset \mathbb{C}$ be open and connected, and suppose $A(\sigma) : X \to Y$, $\sigma \in \Omega$, is an analytic family of Fredholm operators. Then either $A(\sigma)$ is not invertible for any $\sigma \in \Omega$ or $A(\sigma)^{-1}$ is finite-meromorphic.

See Hintz 2025 Theorem 5.77.

Original Quotations

⟨…⟩ marks a scribal insertion (interlinear or marginal); ⟦…⟧ a deletion, with the cancelled text given where legible and *illeg.* where the stroke has rendered it unrecoverable. Spelling, accentuation and punctuation follow the manuscript.

Chapter 2

[1] Pour nous il nous parait fort important de ne donner au principe de la communication de la chaleur aucune extension hypothétique et nous pensons ⟦*illeg.*⟧ que ce principe suffit pour établir la théorie mathématique. ⟨Car⟩ ⟦*illeg.*⟧ les équations sont démontrées ⟦de la maniere la plus claire et la plus rigoureuse⟧ ⟨analytiq.$^{\underline{\text{mt}}}$⟩ sans que l'on ait à examiner si la propagation s'exerce par voye de rayonnement dans l'intérieur ⟨des solides⟩, ⟦*illeg.*⟧ si elle consiste ou non dans l'émission d'une matiere propre que les molecules s'envoyent réciproquement, ou si ⟦*illeg.*⟧ elle resulte comme ⟦*illeg.*⟧ ⟨le son⟩ des vibrations d'un milieu élastique. Il est toujours préférable de se borner a ⟦*illeg.*⟧ ⟨énoncer⟩ le fait général que les observations indiquent, qui n'est autre que le principe précédent. on montre ⟨ainsi⟩ ⟦*illeg.*⟧ que la théorie mathématique de la chaleur est indépendante de toute hypothèse physique; en effet les lois aux quelles la propagation est assujettie sont admises de tous les physiciens malgré l'extreme diversité de leurs sentimens sur la nature et le mode de son action. [p. 35]

[2] nous n'avons point parlé jusqu'ici de la nature physique de la chaleur car ⟦*illeg.*⟧ cette connoissance n'est point nécessaire pour fonder la théorie mathématique. mais on ne peut lier entre eux tous les phénomènes sans se représenter sous une forme distincte ⟦la nature⟧ et l'action du principe. on commettroit une grande erreur si l'on comparoit la chaleur à un fluide élastique comme l'air ou les gaz. la matière est ⟦*illeg.*⟧ dans le monde sous quatre formes différentes ⟦*illeg.*⟧ les substances solides les masses liquides les fluides aériformes ou gaz les substances rayonnantes comme la lumière qui traversent avec une vitesse immense ⟦*illeg.*⟧ tous les points de l'espace en sorte qu'il n'y a aucun point qui ne soit le centre d'une sphère de lumière. ce mode de l'être universel est celui qui convient au principe de la chaleur. [p. 36]

[3] M Biot ne connaissait point alors cette expression, et la remplaçait par une differentielle, c'est à dire qu'il representait par un terme infiniment petit un elem$^{\underline{\text{t}}}$ physique dont la grandeur est finie. … or celle qui s'écoule par toute la superficie n'est point mesurée par une quantité differentielle, donc le facteur qui doit multiplier dt dans l'expression de la chaleur transmise a une valeur finie. on déduit rigoureusement cette expression du seul principe proposé par Newton et Lambert. [p. 40]

[4] le meme auteur assure qu'on ne peut lever la difficulté qui l'arretait et l'empechait de former une équation homogène, qu'en supposant que la chaleur rayonne dans l'intérieur des solides. nous voyons au contraire que cette difficulté disparait entierement aussitôt que l'on compare l'état d'une tranche du prisme à celui d'un solide compris entre deux masses parallèles et inegalement échauffées. ⟦pour juger que la quantité de chaleur qui traverse pendant un certain temps une section du prisme n'est point infiniment petite⟧ ⟨et pour en trouver l'expression exacte⟩ on n'a besoin d'aucune considération physique

sur la maniere dont la Chaleur se propage. il suffit de voir qu'un solide ne conserve son état que lorsqu'il reçoit autant de Chaleur qu'il en perd. [p. 40]

[5] ⟨la géométrie suppose la notion de l'espace et de ses parties. Le ⟦*illeg.*⟧ volume est une partie de⟩ l'espace ⟦*illeg.*⟧ un volume donné étant divisé en deux parties, ce qui est commun à ces deux parties est une surface (les deux parties ⟦hors⟧ dites contiguës) une surface étant divisée en deux parties ce qui est commun aux deux parties contiguës est une ligne. la ligne étant divisée en parties ce qui est commun à deux parties dites contiguës est un point. ⟨⟦*illeg.*⟧⟩ [p. 75]

Chapter 3

[1] ⟨la figure de l'un⟩ De ces polygones approche d'autant plus d'être celle de la ligne courbe que le nombre de côtés est plus grand. or on va reconnaitre des propriétés qui appartiennent à l'une quelconque de ces polygones et on en conclurra qu'elles appartiennent aussi à la courbe. [p. 93]

Bibliography

Primary Sources

Aristotle

1984 *The Complete Works of Aristotle. Bollingen Series LXXI*, ed. by Jonathan Barnes, Electronic edition, Princeton, New Jersey, U.S.A.: Princeton University Press and The Jowett Copyright Trustees, ISBN: 978-1-57085-003-5.

Comte, Auguste

1892 *Cours de philosophie positive. Tome premier contenant les préliminaires généraux et la philosophie mathématique*, 5th, Fifth edition, identical to the first edition published in July 1830., Paris: Au Siège de la Société Positiviste, https://archive.org/details/coursdephilosoph01comt.

De l'Hôpital, Guillaume François Antoine

1696 *Analyse des infiniment petits, pour l'intelligence des lignes courbes*, trans. by Robert E. Bradley, Salvatore J. Petrilli, and C. Edward Sandifer, Science Networks. Historical Studies, 50, Published as *L'Hôpital's Analyse des infiniments petits: An Annotated Translation with Source Material by Johann Bernoulli*, Cham: Springer, ISBN: 978-3-319-17114-2; trans. Paris 1696.

Dirichlet, Peter Gustav Lejeune

1829 Sur la convergence des séries trigonométriques qui servent à représenter une fonction arbitraire entre des limites données, *Journal für die reine und angewandte Mathematik*, 4, pp. 157-169, DOI: 10.48550/arXiv.0806.1294.

Ehrenfest, Paul and Tatjana Ehrenfest

1907 Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem, *Physikalische Zeitschrift*, 8, pp. 311-314, https://archive.org/details/bub_gb_Vy0KAAAAIAAJ/page/n337/mode/2up.

Fourier, Joseph

1808 "Lettre II", in Fourier (1980), pp. 24-26, BnF MS fr. 22501, fol. 68; English translation as letter XX in Herivel 1975, pp. 316-318.

1810a "Document III", in Fourier (1980), pp. 27-35, BnF MS fr. 22501, fols. 76–81; English translation as letter XIX in Herivel 1975, pp. 307-315.

1810b "Lettre IV", in Fourier (1980), pp. 36-38, English translation as letter XVII in Herivel 1975, pp. 302-304; the translation is seriously defective.

1822 *Théorie analytique de la chaleur*, Paris: Firmin Didot.

1888 *Oeuvres de Fourier: Publiées par les soins de Gaston Darboux*, ed. by Jean Gaston Darboux, Cambridge University Press, ISBN: 9781139568081.

Fourier, Joseph

1980 *Joseph Fourier face aux objections contre sa théorie de la chaleur. Lettres inédites, 1808–1816*, ed. by John Herivel, in collab. with Pierre Costabel, Mémoires de la Section des sciences, 8, Paris: Bibliothèque nationale, 88 pp.

n.d.(a) "BnF MS fr. 22519: Recherches en géométrie", 73 fols., Paris, Described in Charbonneau, Catalogue, pp. 103-105. Bibliothèque nationale de France, Département des Manuscrits.

n.d.(b) "BnF MS fr. 22525: Théorie de la chaleur III. Histoire de la théorie. Observations sur les mémoires de Poisson", 219 fols., Paris, Described in Charbonneau, Catalogue, pp. 129-133. Bibliothèque nationale de France, Département des Manuscrits.

Fourier, Joseph, André-Marie Ampère, and François Arago

1822 Rapport fait à l'Académie sur un Mémoire de M. Fresnel relatif à la double réfraction, *Annales de chimie et de physique*, 2nd ser., 20, Reprinted in *Œuvres complètes d'Augustin Fresnel*, vol. 2 (1868), 459–464., pp. 337-344, https://books.google.com/books?id=g6tzUG7JmoQC.

Hamilton, William Rowan

1843 On Fluctuating Functions, *Transactions of the Royal Irish Academy*, 19, Read June 22, 1840., pp. 264-321, https://www.maths.tcd.ie/pub/HistMath/People/Hamilton/Fluctuating/Fluct.pdf.

Laplace, Pierre-Simon

1779 "Mémoire sur les suites", in, *Œuvres complètes de Laplace*, Paris: Gauthier-Villars, vol. 10, pp. 1-89.

1782 "Théorie des attractions des sphéroïdes et de la figure des planètes", in, *Œuvres complètes de Laplace*, Paris: Gauthier-Villars, vol. 10, pp. 339-419.

1799 "Traité de mécanique céleste", in, *Œuvres complètes de Laplace*, Paris: Gauthier-Villars, vol. 2.

1810a "Mémoire sur les intégrales définies et leur application aux probabilités, et spécialement à la recherche du milieu qu'il faut choisir entre les résultats des observations", in, *Œuvres complètes de Laplace*, Paris: Gauthier-Villars, vol. 12, pp. 357-412.

1810b "Mémoire sur les mouvements de la lumière dans les milieux diaphanes", in, *Œuvres complètes de Laplace*, Paris: Gauthier-Villars, vol. 12, pp. 267-298.

1812 "Théorie analytique des probabilités", in, *Œuvres complètes de Laplace*, 3rd ed., Paris: Gauthier-Villars, vol. 7.

Leibniz, Gottfried Wilhelm

1675 "Analyseos tetragonisticae pars secunda", in, *Sämtliche Schriften und Briefe*, ed. by Uwe Mayer, Siegmund Probst, and Heike Sefrin-Weis, Siebente Reihe: Mathematische Schriften, Berlin: Akademie Verlag, vol. 5.1674–1676: Infinitesimalmathematik, pp. 288-295.

Newton, Isaac

1952 *Opticks: Or, A Treatise of the Reflections, Refractions, Inflections, and Colours of Light*, Based on the fourth edition, London, 1730., New York: Dover Publications, ISBN: 9780486602059.

1999 *The Principia: Mathematical Principles of Natural Philosophy*, trans. by I. Bernard Cohen and Anne Whitman, Berkeley, CA: University of California Press, ISBN: 9780520290884.

Poisson, Siméon Denis

1815 Mémoire sur la théorie de la chaleur, *Journal de l'École Polytechnique*, 12, 19, pp. 1-144.

1835 *Théorie mathématique de la chaleur*, Paris: Bachelier.

Prévost, Pierre

1832 *Exposition élémentaire des principes qui servent de base à la théorie de la chaleur rayonnante, faisant suite à l'ouvrage intitulé Du calorique rayonnant*, Geneva and Paris: Abraham Cherbuliez, 149 pp., https://books.google.nl/books?id=CAo_AAAAYAAJ&pg=PA93.

Secondary Sources

Arnold, D. H.

1983 The Mécanique Physique of Siméon Denis Poisson: The Evolution and Isolation in France of his Approach to Physical Theory (1800–1840). IV. Disquiet with Respect to Fourier's Treatment of Heat, *Archive for History of Exact Sciences*, 28, pp. 299-32, doi: 10.1007/BF00328041.

Bos, Henk J. M.

1974 Differentials, Higher-Order Differentials and the Derivative in the Leibnizian Calculus, *Archive for History of Exact Sciences*, 14, 1, pp. 1-90, doi: 10.1007/BF00327456.

Chang, Hasok

2004 *Inventing Temperature: Measurement and Scientific Progress*, Oxford University Press, isbn: 9780199835003.

Charbonneau, Louis

1994 *Catalogue des manuscrits de Joseph Fourier conservés au Cabinet des manuscrits de la Bibliothèque nationale. Fonds français 22501 à 22529*, Paris: Bibliothèque nationale.

Cohen, I. Bernard

1999 A Guide to Newton's Principia, in *The Principia: Mathematical Principles of Natural Philosophy*, Primary text authored by Isaac Newton, translated by I. Bernard Cohen and Anne Whitman., Berkeley, CA: University of California Press, isbn: 9780520290884.

Dhombres, Jean and Jean-Bernard Robert

1998 *Joseph Fourier 1768–1830: Créateur de la physique mathématique*, Paris: Belin, isbn: 2701112133.

Eşanu, Andreea

2019 Auguste Comte and J. S. Mill on Physical Causes: The Case of Joseph Fourier's Analytical Theory of Heat, *HOPOS: The Journal of the International Society for the History of Philosophy of Science*, 9, 2, pp. 275-295, doi: 10.1086/704373.

Fox, Robert

1971 *The Caloric Theory of Gases: From Lavoisier to Regnault*, Clarendon Press, isbn: 9780198581314.

1974 The Rise and Fall of Laplacian Physics, *Historical Studies in the Physical Sciences*, 4, pp. 89-136, doi: 10.2307/27757328.

Friedman, Robert Marc

1977 The Creation of a New Science: Joseph Fourier's Analytical Theory of Heat, *Historical Studies in the Physical Sciences*, 8, pp. 73-99, doi: 10.2307/27757368.

Gibson, George A.

1892 On the History of the Fourier Series, *Proceedings of the Edinburgh Mathematical Society*, 11, pp. 137-166, doi: 10.1017/S001309150003131X.

Grattan-Guinness, Ivor

1990 *Convolutions in French Mathematics, 1800–1840*, 3 vols., Basel: Birkhäuser, isbn: 9783034878111.

Grattan-Guinness, Ivor and J. R. Ravetz

1972 *Joseph Fourier, 1768-1830*, Cambridge, MA: The MIT Press, isbn: 9780262070416.

Graves, Robert Perceval

1882 *Life of Sir William Rowan Hamilton*, Dublin: Hodges, Figgis, & Co., vol. 1, https://archive.org/details/lifeofsirwilliam01gravuoft (visited on 07/23/2026).

Herivel, John

1966 Aspects of French Theoretical Physics in the Nineteenth Century, *The British Journal for the History of Science*, 3, 2, pp. 109-132, doi: 10.1017/S0007087400003794.

1975 *Joseph Fourier: The Man and the Physicist*, Oxford: Clarendon Press, isbn: O198581491.

Jahnke, Hans Niels

2003 (ed.), *A History of Analysis*, History of Mathematics, Providence, RI: American Mathematical Society, vol. 24, isbn: 9780821826232.

Smith, George E.

2002 The Methodology of the *Principia*, in *The Cambridge Companion to Newton*, ed. by I. Bernard Cohen and George E. Smith, Cambridge: Cambridge University Press, pp. 138-173, doi: 10.1017/CCO9781139058568.006.

Stedall, Jacqueline

2008 *Mathematics Emerging. A Sourcebook 1540–1900*, Oxford: Oxford University Press, isbn: 978-0-19-922690-0.

Todhunter, Isaac

1873 *A history of the mathematical theories of attraction and the figure of the earth from the time of Newton to that of Laplace. Volume 2*, London: Macmillan and Co., https://www.gutenberg.org/ebooks/69832.

Truesdell, Clifford

1980 *The Tragicomical History of Thermodynamics, 1822–1854*, Studies in the History of Mathematics and Physical Sciences, New York: Springer-Verlag, vol. 4, doi: 10.1007/978-1-4613-9444-0.

Mathematical Sources

Bertrand, Jacqueline, Pierre Bertrand, and Jean-Philippe Ovarlez

2000 The Mellin Transform, in *The Transforms and Applications Handbook*, ed. by Alexander D. Poularikas, 2nd, Boca Raton: CRC Press, isbn: 9780849383427.

Biswas, Sougata and Jiten C. Kalita

2017 HOC Simulation of Moffatt Eddies and Its Flow Topology in the Triangular Cavity Flow, *arXiv preprint arXiv:1710.06251*, DOI: 10.48550/arXiv.1710.06251.

Branicki, M. and H. K. Moffatt

2006 Evolving Eddy Structures in Oscillatory Stokes Flows in Domains with Sharp Corners, *Journal of Fluid Mechanics*, 551, pp. 63-92, DOI: 10.1017/S002211200500827X.

Conway, John B.

2007 *A Course in Functional Analysis*, 2nd ed., Graduate Texts in Mathematics, New York: Springer, vol. 96, ISBN: 978-0-387-97245-9, DOI: 10.1007/978-1-4757-4383-8.

Flajolet, Philippe and Robert Sedgewick

2009 *Analytic Combinatorics*, Cambridge: Cambridge University Press, ISBN: 9780511801655.

Hintz, Peter

2023 *Lectures on Geometric Singular Analysis, with Applications to Elliptic and Hyperbolic PDE*, Lecture notes, https://bpb-us-e1.wpmucdn.com/sites.psu.edu/dist/3/198411/files/2025/07/unc-pde.pdf.

2025 *An Introduction to Microlocal Analysis*, 1st ed., Graduate Texts in Mathematics, Cham: Springer, ISBN: 9783031907067.

Lesch, Matthias

1997 *Differential Operators of Fuchs Type, Conical Singularities, and Asymptotic Methods*, Teubner-Texte zur Mathematik, Stuttgart-Leipzig: B. G. Teubner Verlag, vol. 136, arXiv: dg-ga/9607005.

Paris, R. B. and D. Kaminski

2001 *Asymptotics and Mellin–Barnes Integrals*, Encyclopedia of Mathematics and its Applications, Cambridge: Cambridge University Press, vol. 85, ISBN: 9780521790017.

Schulze, Bert-Wolfgang

1991 *Pseudo-Differential Operators on Manifolds with Singularities*, Studies in Mathematics and Its Applications, Elsevier, vol. 24, ISBN: 9780080875453.

Vretblad, Anders

2003 *Fourier Analysis and Its Applications*, Graduate Texts in Mathematics, New York: Springer, vol. 223, ISBN: 9780387217239.