LECTURE NOTES ON APPLIED MATHEMATICS

Methods and Models

John K. Hunter
Department of Mathematics
University of California, Davis

June 17, 2009

Contents

LECTURE 1

Introduction

> The source of all great mathematics is the special case, the concrete example. It is frequent in mathematics that every instance of a concept of seemingly great generality is in essence the same as a small and concrete special case.[1]

We begin by describing a rather general framework for the derivation of PDEs that describe the conservation, or balance, of some quantity.

1. Conservation laws

We consider a quantity $\mathcal{Q}$ that varies in space, $\vec{x}$, and time, t, with density $u(\vec{x},t)$, flux $\vec{q}(\vec{x},t)$, and source density $\sigma(\vec{x},t)$.

For example, if $\mathcal{Q}$ is the mass of a chemical species diffusing through a stationary medium, we may take u to be the density, $\vec{q}$ the mass flux, and f the mass rate per unit volume at which the species is generated.

For simplicity, we suppose that $u(x,t)$ is scalar-valued, but exactly the same considerations would apply to a vector-valued density (leading to a system of equations).

1.1. Integral form

The conservation of $\mathcal{Q}$ is expressed by the condition that, for any fixed spatial region Ω, we have

$$\frac{d}{dt}\int_\Omega u\,d\vec{x} = -\int_{\partial\Omega}\vec{q}\cdot\vec{n}\,dS + \int_\Omega \sigma\,d\vec{x}. \tag{1.1}$$

Here, $\partial\Omega$ is the boundary of Ω, $\vec{n}$ is the unit outward normal, and dS denotes integration with respect to surface area.

Equation (1.1) is the integral form of conservation of $\mathcal{Q}$. It states that, for any region Ω, the rate of change of the total amount of $\mathcal{Q}$ in Ω is equal to the rate at which $\mathcal{Q}$ flows into Ω through the boundary $\partial\Omega$ plus the rate at which $\mathcal{Q}$ is generated by sources inside Ω.

1.2. Differential form

Bringing the time derivative in (1.1) inside the integral over the fixed region Ω, and using the divergence theorem, we may write (1.1) as

$$\int_\Omega u_t\,d\vec{x} = \int_\Omega \left(-\nabla\cdot\vec{q}+\sigma\right)\,d\vec{x}$$

[1] P. Halmos.

Since this equation holds for arbitrary regions Ω, it follows that, for smooth functions,

$$u_t = -\nabla \cdot \vec{q} + \sigma. \tag{1.2}$$

Equation (1.2) is the differential form of conservation of $\mathcal{Q}$.

When the source term σ is nonzero, (1.2) is often called, with more accuracy, a balance law for $\mathcal{Q}$, rather than a conservation law, but we won't insist on this distinction.

2. Constitutive equations

The conservation law (1.2) is not a closed equation for the density u. Typically, we supplement it with constitutive equations that relate the flux $\vec{q}$ and the source density σ to u and its derivatives. While the conservation law expresses a general physical principle, constitutive equations describe the response of a particular system being modeled.

Example 1.1. If the flux and source are pointwise functions of the density,

$$\vec{q} = \vec{f}(u), \qquad \sigma = g(u),$$

then we get a first-order system of PDEs

$$u_t + \nabla \cdot \vec{f}(u) = g(u).$$

For example, in one space dimension, if $g(u) = 0$ and $f(u) = u^2/2$, we get the inviscid Burgers equation

$$u_t + \left(\frac{1}{2}u^2\right)_x = 0.$$

This equation is a basic model equation for hyperbolic systems of conservation laws, such as the compressible Euler equations for the flow of an inviscid compressible fluid [**47**].

Example 1.2. Suppose that the flux is a linear function of the density gradient,

$$\vec{q} = -A\nabla u, \tag{1.3}$$

where A is a second-order tensor, that is a linear map between vectors. It is represented by an $n \times n$ matrix with respect to a choice of n basis vectors. Then, if $\sigma = 0$, we get a second order, linear PDE for $u(\vec{x}, t)$

$$u_t = \nabla \cdot (A\nabla u). \tag{1.4}$$

Examples of this constitutive equation include: Fourier's law in heat conduction (heat flux is a linear function of temperature gradient); Fick's law (flux of solute is a linear function of the concentration gradient); and Darcy's law (fluid velocity in a porous medium is a linear function of the pressure gradient). It is interesting to note how old each of these laws is: Fourier (1822); Fick (1855); Darcy (1855).

The conductivity tensor A in (1.3) is usually symmetric and positive-definite, in which case (1.4) is a parabolic PDE; the corresponding PDE for equilibrium density distributions $u(\vec{x})$ is then an elliptic equation

$$\nabla \cdot (A\nabla u) = 0.$$

In general, the conductivity tensor may depend upon $\vec{x}$ in a nonuniform system, and on u in non-linearly diffusive systems. While A is almost always symmetric,

it need not be diagonal in an anisotropic system. For example, the heat flux in a crystal lattice or in a composite medium made up of alternating thin layers of copper and asbestos is not necessarily in the same direction as the temperature gradient.

For a uniform, isotropic, linear system, we have $A = \nu I$ where ν is a positive constant, and then $u(\vec{x}, t)$ satisfies the heat, or diffusion, equation

$$u_t = \nu \Delta u.$$

Equilibrium solutions satisfy Laplace's equation

$$\Delta u = 0.$$

3. The KPP equation

In this section, we discuss a specific example of an equation that arises as a model in population dynamics and genetics.

3.1. Reaction-diffusion equations

If $\vec{q} = -\nu \nabla u$ and $\sigma = f(u)$ in (1.2), we get a *reaction-diffusion* equation

$$u_t = \nu \Delta u + f(u).$$

Spatially uniform solutions satisfy the ODE

$$u_t = f(u),$$

which is the 'reaction' equation. In addition, diffusion couples together the solution at different points.

Such equations arise, for example, as models of spatially nonuniform chemical reactions, and of population dynamics in spatially distributed species.

The combined effects of spatial diffusion and nonlinear reaction can lead to the formation of many different types of spatial patterns; the spiral waves that occur in Belousov-Zabotinski reactions are one example.

One of the simplest reaction-diffusion equations is the KPP equation (or Fisher equation)

$$u_t = \nu u_{xx} + ku(a - u). \tag{1.5}$$

Here, ν, k, a are positive constants; as we will show, they may be set equal to 1 without loss of generality.

Equation (1.5) was introduced independently by Fisher [**22**], and Kolmogorov, Petrovsky, and Piskunov [**33**] in 1937. It provides a simple model for the dispersion of a spatially distributed species with population density $u(x, t)$ or, in Fisher's work, for the advance of a favorable allele through a spatially distributed population.

3.2. Maximum principle

According to the maximum principle, the solution of (1.5) remains nonnegative if the initial data $u_0(x) = u(x, 0)$ is non-negative, which is consistent with its use as a model of population or probability.

The maximum principle holds because if u first crosses from positive to negative values at time t_0 at the point x_0, and if $u(x, t)$ has a nondegenerate minimum at x_0, then $u_{xx}(x_0, t_0) > 0$. Hence, from (1.5), $u_t(x_0, t_0) > 0$, so u cannot evolve forward in time into the region $u < 0$. A more careful argument is required to deal with degenerate minima, and with boundaries, but the conclusion is the same [**18, 42**].

A similar argument shows that $u(x, t) \leq 1$ for all $t \geq 0$ if $u_0(x) \leq 1$.

Remark 1.3. A forth-order diffusion equation, such as

$$u_t = -u_{xxxx} + u(1-u),$$

does not satisfy a maximum principle, and it is possible for positive initial data to evolve into negative values.

3.3. Logistic equation

Spatially uniform solutions of (1.5) satisfy the logistic equation

$$u_t = ku(a-u). \tag{1.6}$$

This ODE has two equilibrium solutions at $u = 0$, $u = a$.

The solution $u = 0$ corresponds to a complete absence of the species, and is unstable. Small disturbances grow initially like $u_0 e^{kat}$. The solution $u = a$ corresponds to the maximum population that can be sustained by the available resources. It is globally asymptotically stable, meaning that any solution of (1.6) with a strictly positive initial value approaches a as $t \to \infty$.

Thus, the PDE (1.5) describes the evolution of a population that satisfies logistic dynamics at each point of space coupled with dispersal into regions of lower population.

3.4. Nondimensionalization

Before discussing (1.5) further, we simplify the equation by rescaling the variables to remove the constants. Let

$$u = U\bar{u}, \qquad x = L\bar{x}, \qquad t = T\bar{t}$$

where U, L, T are arbitrary positive constants. Then

$$\frac{\partial}{\partial x} = \frac{1}{L}\frac{\partial}{\partial \bar{x}}, \frac{\partial}{\partial t} = \frac{1}{T}\frac{\partial}{\partial \bar{t}}.$$

It follows that $\bar{u}\,(\bar{x},\bar{t})$ satisfies

$$\bar{u}_{\bar{t}} = \left(\frac{\nu T}{L^2}\right)\bar{u}_{\bar{x}\bar{x}} + (kTU)\,\bar{u}\left(\frac{a}{U} - \bar{u}\right).$$

Therefore, choosing

$$U = a, \qquad T = \frac{1}{ka}, \qquad L = \sqrt{\frac{\nu}{ka}}, \tag{1.7}$$

and dropping the bars, we find that $u(x,t)$ satisfies

$$u_t = u_{xx} + u(1-u). \tag{1.8}$$

Thus, in the absence of any other parameters, none of the coefficients in (1.5) are essential.

If we consider (1.5) on a finite domain of length ℓ, then the problem depends in an essential way on a dimensionless constant R, which we may write as

$$\mathrm{R} = \frac{ka\ell^2}{\nu}.$$

We could equivalently use $1/\mathrm{R}$ or $\sqrt{\mathrm{R}}$, or some other expression, instead of R. From (1.7), we have $\mathrm{R} = T_d/T_r$ where $T_r = T$ is a timescale for solutions of the reaction equation (1.6) to approach the equilibrium value a, and $T_d = \ell^2/\nu$ is a timescale for linear diffusion to significantly influence the entire length ℓ of the domain. The qualitative behavior of solutions depends on R.

When dimensionless parameters exist, we have a choice in how we define dimensionless variables. For example, on a finite domain, we could nondimensionalize as above, which would give (1.8) on a domain of length $\sqrt{R}$. Alternatively, we might prefer to use the length ℓ of the domain to nondimensionalize lengths. In that case, the nondimensionalized domain has length 1, and the nondimensionalized form of (1.5) is

$$u_t = \frac{1}{\mathrm{R}} u_{xx} + u\,(1-u)\,.$$

We get a small, or large, dimensionless diffusivity if the diffusive timescale is large, or small, respectively, compared with the reaction time scale.

Somewhat less obviously, even on infinite domains additional lengthscales may be introduced into a problem by initial data

$$u(x,0) = u_0(x).$$

Using the variables (1.7), we get the nondimensionalized initial condition

$$\bar{u}\,(\bar{x},0) = \bar{u}_0\,(\bar{x})\,,$$

where

$$\bar{u}_0\,(\bar{x}) = \frac{1}{a} u_0\,(L\bar{x})\,.$$

Thus, for example, if u_0 has a typical amplitude a and varies over a typical lengthscale of ℓ, then we may write

$$u_0(x) = a\bar{f}\left(\frac{x}{\ell}\right)$$

where $\bar{f}$ is a dimensionless function. Then

$$\bar{u}_0\,(\bar{x}) = \bar{f}\left(\sqrt{\mathrm{R}}\bar{x}\right),$$

and the evolution of the solution depends upon whether the initial data varies rapidly, slowly, or on the same scale as the reaction-diffusion length scale L.

3.5. Traveling waves

One of the principal features of the KPP equation is the existence of traveling waves which describe the invasion of an unpopulated region (or a region whose population does not possess the favorable allele) from an adjacent populated region.

A traveling wave is a solution of the form

$$u(x,t) = f(x-ct) \tag{1.9}$$

where c is a constant wave speed. This solution consists of a fixed spatial profile that propagates with velocity c without changing its shape.

For definiteness we assume that $c > 0$. The case $c < 0$ can be reduced to this one by a reflection $x \mapsto -x$, which transforms a right-moving wave into a left-moving wave.

Use of (1.9) in (1.8) implies that $f(x)$ satisfies the ODE

$$f'' + cf' + f(1-f) = 0. \tag{1.10}$$

The equilibria of this ODE are $f = 0$, $f = 1$.

Note that (1.10) describes the spatial dynamics of traveling waves, whereas (1.6) describes the temporal dynamics of uniform solutions. Although these equations have the same equilibrium solutions, they are different ODEs (for example, one

is second order, and the other first order) and the stability of their equilibrium solutions means different things.

The linearization of (1.10) at $f = 0$ is

$$f'' + cf' + f = 0.$$

The characteristic equation of this ODE is

$$\lambda^2 + c\lambda + 1 = 0$$

with roots

$$\lambda = \frac{1}{2}\left\{-c \pm \sqrt{c^2 - 4}\right\}.$$

Thus, the equilibrium $f = 0$ is a stable spiral point if $0 < c < 2$, a degenerate stable node if $c = 2$, and a stable node if $2 < c < \infty$.

The linearization of (1.10) at $f = 1$ is

$$f'' + cf' - f = 0.$$

The characteristic equation of this ODE is

$$\lambda^2 + c\lambda - 1 = 0$$

with roots

$$\lambda = \frac{1}{2}\left\{-c \pm \sqrt{c^2 + 4}\right\}.$$

Thus, the equilibrium $f = 1$ is a saddlepoint.

As we will show next, for any $2 \leq c < \infty$ there is a unique positive heteroclinic orbit $F(x)$ connecting the unstable saddle point at $f = 1$ to the stable equilibrium at $f = 0$, meaning that

$$F(x) \to 1 \quad \text{as } x \to -\infty; \qquad F(x) \to 0 \quad \text{as } x \to \infty.$$

These right-moving waves describe the invasion of the state $u = 0$ by the state $u = 1$. Reflecting $x \mapsto -x$, we get a corresponding family of left-moving traveling waves with $-\infty < c \leq -2$.

Since the traveling wave ODE (1.10) is autonomous, if $F(x)$ is a solution then so is $F(x - x_0)$ for any constant x_0. This solution has the same orbit as $F(x)$, and corresponds to a traveling wave of the same velocity that is translated by a constant distance x_0.

There is also a traveling wave solution for $0 < c < 2$ However, in that case the solution becomes negative near 0 since $f = 0$ is a spiral point. This solution is therefore not relevant to the biological application we have and mind. Moreover, by the maximum principle, it cannot arise from nonnegative initial data.

The traveling wave most relevant to the applications considered above is, perhaps, the positive one with the slowest speed ($c = 2$); this is the one that describes the mechanism of diffusion from the populated region into the unpopulated one, followed by logistic growth of the diffusive perturbation. The faster waves arise because of the growth of small, but nonzero, pre-existing perturbations of the unstable state $u = 0$ ahead of the wavefront.

The linear instability of the state $u = 0$ is arguably a defect of the model. If there were a threshold below which a small population died out, then this dependence of the wave speed on the decay rate of the initial data would not arise.

3.6. The existence of traveling waves

Let us discuss the existence of positive traveling waves in a little more detail.

If $c = 5/\sqrt{6}$, there is a simple explicit solution for the traveling wave [**1**]:

$$F(x) = \frac{1}{\left(1 + e^{x/\sqrt{6}}\right)^2}.$$

Although there is no similar explicit solution for general values of c, we can show the existence of traveling waves by a qualitative argument.

Writing (1.10) as a first order system of ODEs for (f, g), where $g = f'$, we get

$$\begin{aligned} f' &= g, \\ g' &= -f(1-f) - cg. \end{aligned} \tag{1.11}$$

For $c \geq 2$, we choose $0 < \beta \leq 1$ such that

$$\beta + \frac{1}{\beta} = c, \qquad \beta = \frac{1}{2}\left(c - \sqrt{c^2 - 4}\right).$$

Then, on the line $g = -\beta f$ with $0 < f \leq 1$, the trajectories of the system satisfy

$$\frac{dg}{df} = \frac{g'}{f'} = -c - \frac{f(1-f)}{g} = -c + \frac{1-f}{\beta} < -c + \frac{1}{\beta} = -\beta.$$

Since $f' < 0$ for $g < 0$, and $dg/df < -\beta$, the trajectories of the ODE enter the triangular region

$$D = \{(f, g) : 0 < f < 1, -\beta f < g < 0\}.$$

Moreover, since $g' < 0$ on $g = 0$ when $0 < f < 1$, and $f' < 0$ on $f = 1$ when $g < 0$, the region D is positively invariant (meaning that any trajectory that starts in the region remains in the region for all later times).

The linearization of the system (1.11) at the fixed point $(f, g) = (1, 0)$ is

$$\begin{pmatrix} f' \\ g' \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -c \end{pmatrix} \begin{pmatrix} f \\ g \end{pmatrix}.$$

The unstable manifold of $(1, 0)$, with corresponding eigenvalue

$$\lambda = \frac{1}{2}\left(-c + \sqrt{c^2 + 4}\right) > 0,$$

is in the direction

$$\vec{r} = \begin{pmatrix} -1 \\ -\lambda \end{pmatrix}.$$

The corresponding trajectory below the f-axis must remain in D, and since D contains no other fixed points or limit cycles, it must approach the fixed point $(0, 0)$ as $x \to \infty$.

Thus, a nonnegative traveling wave connecting $f = 1$ to $f = 0$ exists for every $c \geq 2$.

3.7. The initial value problem

Consider the following initial value problem for the KPP equation

$$\begin{aligned} &u_t = u_{xx} + u(1-u), \\ &u(x, 0) = u_0(x), \\ &u(x, t) \to 1 \text{ as } x \to -\infty, \\ &u(x, t) \to 0 \text{ as } x \to \infty. \end{aligned}$$

Kolmogorov, Petrovsky and Piskunov proved that if $0 \leq u_0(x) \leq 1$ is any initial data that is exactly equal to 1 for all sufficiently large negative x, and exactly equal to 0 for all sufficiently large positive x, then the solution approaches the traveling wave with $c = 2$ as $t \to \infty$.

This result is sensitive to a change in the spatial decay rate of the initial data into the unstable state $u = 0$. Specifically, suppose that

$$u_0(x) \sim Ce^{-\beta x}$$

as $x \to \infty$, where β is some positive constant (and C is nonzero). If $\beta \geq 1$, then the solution approaches a traveling wave of speed 2; but if $0 < \beta < 1$, meaning that the initial data decays more slowly, then the solution approaches a traveling wave of speed

$$c(\beta) = \beta + \frac{1}{\beta}.$$

This is the wave speed of the traveling wave solution of (1.10) that decays to $f = 0$ at the rate $f \sim Ce^{-\beta x}$.

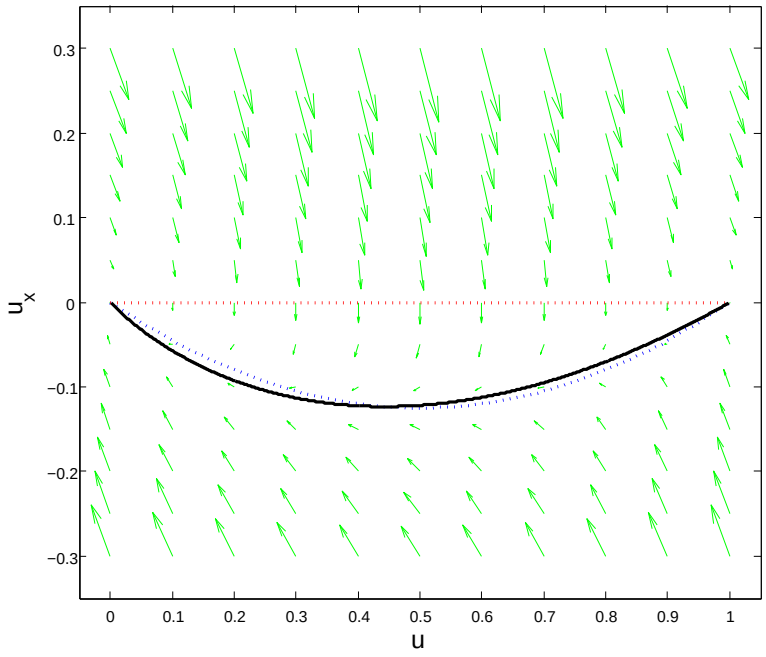

FIGURE 1. The phase plane for the KPP traveling wave, showing the heteroclinic orbit connecting $(1, 0)$ to $(0, 0)$ (courtesy of Tim Lewis).

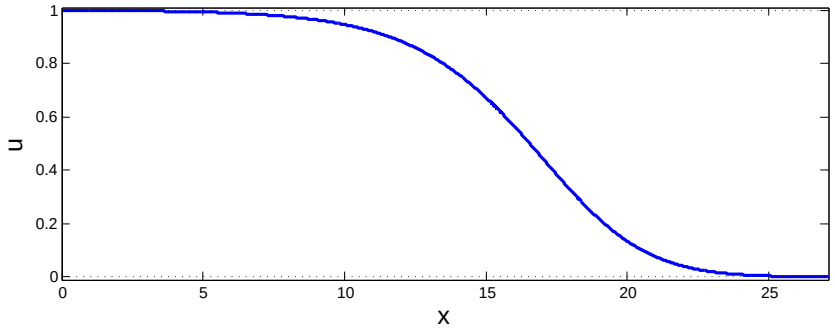

FIGURE 2. The spatial profile of the traveling wave.

LECTURE 2

Dimensional Analysis, Scaling, and Similarity

1. Systems of units

The numerical value of any quantity in a mathematical model is measured with respect to a system of units (for example, meters in a mechanical model, or dollars in a financial model). The units used to measure a quantity are arbitrary, and a change in the system of units (for example, from meters to feet) cannot change the model.

A crucial property of a quantitative system of units is that the value of a dimensional quantity may be measured as some multiple of a basic unit. Thus, a change in the system of units leads to a rescaling of the quantities it measures, and the ratio of two quantities with the same units does not depend on the particular choice of the system. The independence of a model from the system of units used to measure the quantities that appear in it therefore corresponds to a scale-invariance of the model.

Remark 2.1. Sometimes it is convenient to use a logarithmic scale of units instead of a linear scale (such as the Richter scale for earthquake magnitudes, or the stellar magnitude scale for the brightness of stars) but we can convert this to an underlying linear scale. In other cases, qualitative scales are used (such as the Beaufort wind force scale), but these scales ("leaves rustle" or "umbrella use becomes difficult") are not susceptible to a quantitative analysis (unless they are converted in some way into a measurable linear scale). In any event, we will take connection between changes in a system of units and rescaling as a basic premise.

A *fundamental system of units* is a set of independent units from which all other units in the system can be derived. The notion of independent units can be made precise in terms of the rank of a suitable matrix [**7, 10**] but we won't give the details here.

The choice of fundamental units in a particular class of problems is not unique, but, given a fundamental system of units, any other derived unit may be constructed uniquely as a product of powers of the fundamental units.

Example 2.2. In mechanical problems, a fundamental set of units is mass, length, time, or M, L, T, respectively, for short. With this fundamental system, velocity $V = LT^{-1}$ and force $F = MLT^{-2}$ are derived units. We could instead use, say, force F, length L, and time T as a fundamental system of units, and then mass $M = FL^{-1}T^2$ is a derived unit.

Example 2.3. In problems involving heat flow, we may introduce temperature (measured, for example, in Kelvin) as a fundamental unit. The linearity of temperature is somewhat peculiar: although the 'zeroth law' of thermodynamics ensures that equality of temperature is well defined, it does not say how temperatures can

be 'added.' Nevertheless, empirical temperature scales are defined, by convention, to be linear scales between two fixed points, while thermodynamics temperature is an energy, which is additive.

Example 2.4. In problems involving electromagnetism, we may introduce current as a fundamental unit (measured, for example, in Ampères in the SI system) or charge (measured, for example, in electrostatic units in the cgs system). Unfortunately, the officially endorsed SI system is often less convenient for theoretical work than the cgs system, and both systems remain in use.

Not only is the distinction between fundamental and derived units a matter of choice, or convention, the number of fundamental units is also somewhat arbitrary. For example, if dimensional constants are present, we may reduce the number of fundamental units in a given system by setting the dimensional constants equal to fixed dimensionless values.

Example 2.5. In relativistic mechanics, if we use M, L, T as fundamental units, then the speed of light c is a dimensional constant ($c = 3 \times 10^8\,\mathrm{ms}^{-1}$ in SI-units). Instead, we may set $c = 1$ and use M, T (for example) as fundamental units. This means that we measure lengths in terms of the travel-time of light (one nanosecond being a convenient choice for everyday lengths).

2. Scaling

Let $(d_1, d_2, \ldots, d_r)$ denote a fundamental system of units, such as (M, L, T) in mechanics, and a a quantity that is measurable with respect to this system. Then the dimension of a, denoted $[a]$, is given by

$$[a] = d_1^{\alpha_1} d_2^{\alpha_2} \ldots d_r^{\alpha_r} \tag{2.1}$$

for suitable exponents $(\alpha_1, \alpha_2, \ldots, \alpha_r)$.

Suppose that $(a_1, a_2, \ldots, a_n)$ denotes all of the dimensional quantities appearing in a particular model, including parameters, dependent variables, and independent variables. We denote the dimension of a_i by

$$[a_i] = d_1^{\alpha_{1,i}} d_2^{\alpha_{2,i}} \ldots d_r^{\alpha_{r,i}}. \tag{2.2}$$

The invariance of the model under a change in units $d_j \mapsto \lambda_j d_j$ implies that it is invariant under the scaling transformation

$$a_i \to \lambda_1^{\alpha_{1,i}} \lambda_2^{\alpha_{2,i}} \ldots \lambda_r^{\alpha_{r,i}} a_i \qquad i = 1, \ldots, n$$

for any $\lambda_1, \ldots \lambda_r > 0$.

Thus, if

$$a = f(a_1, \ldots, a_n)$$

is any relation between quantities in the model with the dimensions in (2.1) and (2.2), then f must have the scaling property that

$$\lambda_1^{\alpha_1} \lambda_2^{\alpha_2} \ldots \lambda_r^{\alpha_r} f(a_1, \ldots, a_n) = f\left(\lambda_1^{\alpha_{1,1}} \lambda_2^{\alpha_{2,1}} \ldots \lambda_r^{\alpha_{r,1}} a_1, \ldots, \lambda_1^{\alpha_{1,n}} \lambda_2^{\alpha_{2,n}} \ldots \lambda_r^{\alpha_{r,n}} a_n\right).$$

A particular consequence of this invariance is that any two quantities that are equal must have the same dimension (otherwise a change in units would violate the equality). This fact is often useful in finding the dimension of some quantity.

Example 2.6. According to Newton's second law,

$$\text{force} = \text{rate of change of momentum with respect to time.}$$

Thus, if F denotes the dimension of force and P the dimension of momentum, then $F = P/T$. Since $P = MV = ML/T$, we conclude that $F = ML/T^2$ (or mass $\times$ acceleration).

3. Nondimensionalization

Scale-invariance implies that we can reduce the number of quantities appearing in a problem by introducing dimensionless quantities.

Suppose that $(a_1, \ldots, a_r)$ are a set of quantities whose dimensions form a fundamental system of units. We denote the remaining quantities in the model by $(b_1, \ldots, b_m)$, where $r + m = n$. Then, for suitable exponents $(\beta_{1,i}, \ldots, \beta_{r,i})$ determined by the dimensions of $(a_1, \ldots, a_r)$ and b_i, the quantity

$$\Pi_i = \frac{b_i}{a_1^{\beta_{1,i}} \ldots a_r^{\beta_{r,i}}}$$

is dimensionless, meaning that it is invariant under the scaling transformations induced by changes in units.

A dimensionless parameter Π_i can typically be interpreted as the ratio of two quantities of the same dimension appearing in the problem (such as a ratio of lengths, times, diffusivities, and so on). In studying a problem, it is crucial to know the magnitude of the dimensionless parameters on which it depends, and whether they are small, large, or roughly of the order one.

Any dimensional equation

$$a = f(a_1, \ldots, a_r, b_1, \ldots, b_m)$$

is, after rescaling, equivalent to the dimensionless equation

$$\Pi = f(1, \ldots, 1, \Pi_1, \ldots, \Pi_m).$$

Thus, the introduction of dimensionless quantities reduces the number of variables in the problem by the number of fundamental units. This fact is called the 'Buckingham Pi-theorem.' Moreover, any two systems with the same values of dimensionless parameters behave in the same way, up to a rescaling.

4. Fluid mechanics

To illustrate the ideas of dimensional analysis, we describe some applications in fluid mechanics.

Consider the flow of a homogeneous fluid with speed U and length scale L. We restrict our attention to incompressible flows, for which U is much smaller that the speed of sound c_0 in the fluid, meaning that the Mach number

$$M = \frac{U}{c_0}$$

is small. The sound speed in air at standard conditions is $c_0 = 340\,\mathrm{ms}^{-1}$. The incompressibility assumption is typically reasonable when $M \leq 0.2$.

The physical properties of a viscous, incompressible fluid depend upon two dimensional parameters, its mass density ρ_0 and its (dynamic) viscosity μ. The dimension of the density is

$$[\rho_0] = \frac{M}{L^3}.$$

The dimension of the viscosity, which measures the internal friction of the fluid, is given by

$$[\mu] = \frac{M}{LT}. \tag{2.3}$$

To derive this result, we explain how the viscosity arises in the constitutive equation of a Newtonian fluid relating the stress and the strain rate.

4.1. The stress tensor

The stress, or force per unit area, $\vec{t}$ exerted across a surface by fluid on one side of the surface on fluid on the other side is given by

$$\vec{t} = \mathbf{T}\vec{n}$$

where $\mathbf{T}$ is the Cauchy stress tensor and $\vec{n}$ is a unit vector to the surface. It is a fundamental result in continuum mechanics, due to Cauchy, that $\vec{t}$ is a linear function of $\vec{n}$; thus, $\mathbf{T}$ is a second-order tensor [**25**].

The sign of $\vec{n}$ is chosen, by convention, so that if $\vec{n}$ points into fluid on one side A of the surface, and away from fluid on the other side B, then $\mathbf{T}\vec{n}$ is the stress exerted by A on B. A reversal of the sign of $\vec{n}$ gives the equal and opposite stress exerted by B on A.

The stress tensor in a Newtonian fluid has the form

$$\mathbf{T} = -p\mathbf{I} + 2\mu\mathbf{D} \tag{2.4}$$

where p is the fluid pressure, μ is the dynamic viscosity, $\mathbf{I}$ is the identity tensor, and $\mathbf{D}$ is the strain-rate tensor

$$\mathbf{D} = \frac{1}{2}\left(\nabla\vec{u} + \nabla\vec{u}^{\top}\right).$$

Thus, $\mathbf{D}$ is the symmetric part of the velocity gradient $\nabla\vec{u}$.

In components,

$$T_{ij} = -p\delta_{ij} + \mu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$$

where δ_{ij} is the Kronecker-δ,

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Example 2.7. Newton's original definition of viscosity (1687) was for shear flows. The velocity of a shear flow with strain rate σ is given by

$$\vec{u} = \sigma x_2 \vec{e}_1$$

where $\vec{x} = (x_1, x_2, x_3)$ and $\vec{e}_i$ is the unit vector in the i^{th} direction. The velocity gradient and strain-rate tensors are

$$\nabla\vec{u} = \begin{pmatrix} 0 & \sigma & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{D} = \frac{1}{2}\begin{pmatrix} 0 & \sigma & 0 \\ \sigma & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The viscous stress $\vec{t}_v = 2\mu\mathbf{D}\vec{n}$ exerted by the fluid in $x_2 > 0$ on the fluid in $x_2 < 0$ across the surface $x_2 = 0$, with unit normal $\vec{n} = \vec{e}_2$ pointing into the region $x_2 > 0$, is $\vec{t}_v = \sigma\mu\vec{e}_1$. (There is also a normal pressure force $\vec{t}_p = -p\vec{e}_1$.) Thus, the frictional viscous stress exerted by one layer of fluid on another is proportional the strain rate σ and the viscosity μ.

4.2. Viscosity

The dynamic viscosity μ is a constant of proportionality that relates the strain-rate to the viscous stress.

Stress has the dimension of force/area, so

$$[\mathbf{T}] = \frac{ML}{T^2}\frac{1}{L^2} = \frac{M}{LT^2}.$$

The strain-rate has the dimension of a velocity gradient, or velocity/length, so

$$[\mathbf{D}] = \frac{L}{T}\frac{1}{L} = \frac{1}{T}.$$

Since $\mu\mathbf{D}$ has the same dimension as $\mathbf{T}$, we conclude that μ has the dimension in (2.3).

The *kinematic viscosity* ν of the fluid is defined by

$$\nu = \frac{\mu}{\rho_0}.$$

It follows from (2.3) that ν has the dimension of a diffusivity,

$$[\nu] = \frac{L^2}{T}.$$

The kinematic viscosity is a diffusivity of momentum; viscous effects lead to the diffusion of momentum in time T over a length scale of the order $\sqrt{\nu T}$.

The kinematic viscosity of water at standard conditions is approximately $1\,\mathrm{mm}^2/\mathrm{s}$, meaning that viscous effects diffuse fluid momentum in one second over a distance of the order 1 mm. The kinematic viscosity of air at standard conditions is approximately $15\,\mathrm{mm}^2/\mathrm{s}$; it is larger than that of water because of the lower density of air. These values are small on every-day scales. For example, the timescale for viscous diffusion across room of width 10 m is of the order of 6×10^6 s, or about 77 days.

4.3. The Reynolds number

The dimensional parameters that characterize a fluid flow are a typical velocity U and length L, the kinematic viscosity ν, and the fluid density ρ_0. Their dimensions are

$$[U] = \frac{L}{T}, \quad [L] = L, \quad [\nu] = \frac{L^2}{T}, \quad [\rho_0] = \frac{M}{L^3}.$$

We can form a single independent dimensionless parameter from these dimensional parameters, the *Reynolds number*

$$\mathrm{R} = \frac{UL}{\nu}. \tag{2.5}$$

As long as the assumptions of the original incompressible model apply, the behavior of a flow with similar boundary and initial conditions depends only on its Reynolds number.

The inertial term in the Navier-Stokes equation has the order of magnitude

$$\rho_0 \vec{u} \cdot \nabla \vec{u} = O\left(\frac{\rho_0 U^2}{L}\right),$$

while the viscous term has the order of magnitude

$$\mu \Delta \vec{u} = O\left(\frac{\mu U}{L^2}\right).$$

The Reynolds number may therefore be interpreted as a ratio of the magnitudes of the inertial and viscous terms.

The Reynolds number spans a large range of values in naturally occurring flows, from 10^{-20} in the very slow flows of the earth's mantle, to 10^{-5} for the motion of bacteria in a fluid, to 10^6 for air flow past a car traveling at 60 mph, to 10^{10} in some large-scale geophysical flows.

Example 2.8. Consider a sphere of radius L moving through an incompressible fluid with constant speed U. A primary quantity of interest is the total drag force D exerted by the fluid on the sphere. The drag is a function of the parameters on which the problem depends, meaning that

$$D = f(U, L, \rho_0, \nu).$$

The drag D has the dimension of force (ML/T^2), so dimensional analysis implies that

$$D = \rho_0 U^2 L^2 F\left(\frac{UL}{\nu}\right).$$

Thus, the dimensionless drag

$$\frac{D}{\rho_0 U^2 L^2} = F(\mathrm{R}) \tag{2.6}$$

is a function of the Reynolds number (2.5), and dimensional analysis reduces the problem of finding a function f of four variables to finding a function F of one variable.

The function $F(\mathrm{R})$ has a complicated dependence on R which is difficult to determine theoretically, especially for large values of the Reynolds number. Nevertheless, experimental measurements of the drag for a wide variety of values of U, L, ρ_0 and ν agree well with (2.6)

4.4. The Navier-Stokes equations

The flow of an incompressible homogeneous fluid with density ρ_0 and viscosity μ is described by the incompressible Navier-Stokes equations,

$$\begin{aligned} &\rho_0 \left(\vec{u}_t + \vec{u} \cdot \nabla \vec{u}\right) + \nabla p = \mu \Delta \vec{u}, \\ &\nabla \cdot \vec{u} = 0. \end{aligned} \tag{2.7}$$

Here, $\vec{u}\,(\vec{x}, t)$ is the velocity of the fluid, and $p\,(\vec{x}, t)$ is the pressure. The first equation is conservation of momentum, and the second equation is conservation of volume.

Remark 2.9. It remains an open question whether or not the three-dimensional Navier-Stokes equations, with arbitrary smooth initial data and appropriate boundary conditions, have a unique, smooth solution that is defined for all positive times. This is one of the Clay Institute Millenium Prize Problems.

Let U, L be a typical velocity scale and length scale of a fluid flow, and define dimensionless variables by

$$\vec{u}^* = \frac{\vec{u}}{U}, \quad p^* = \frac{p}{\rho U^2}, \quad \vec{x}^* = \frac{\vec{x}}{L}, \quad t^* = \frac{Ut}{L}.$$

Using these expressions in (2.7), and dropping the stars on the dimensionless variables, we get

$$\begin{aligned} &\vec{u}_t + \vec{u}\cdot\nabla\vec{u} + \nabla p = \frac{1}{\mathrm{R}}\Delta\vec{u}, \\ &\nabla\cdot\vec{u} = 0, \end{aligned} \tag{2.8}$$

where R is the Reynolds number defined in (2.5).

4.5. Euler equations

The nondimensionalized equation (2.8) suggests that for flows with high Reynolds number, we may neglect the viscous term on the right hand side of the momentum equation, and approximate the Navier-Stokes equation by the *incompressible Euler equations*

$$\begin{aligned} &\vec{u}_t + \vec{u}\cdot\nabla\vec{u} + \nabla p = 0, \\ &\nabla\cdot\vec{u} = 0. \end{aligned}$$

The Euler equations are difficult to analyze because, like the Navier-Stokes equations, they are nonlinear. Moreover, the approximation of the Navier-Stokes equation by the Euler equations is problematic. High-Reynolds number flows develop complicated small-scale structures (for instance, boundary layers and turbulence) and, as a result, it is not always possible to neglect the second-order spatial derivatives $\Delta\vec{u}$ in the viscous term in comparison with the first-order spatial derivatives $\vec{u}\cdot\nabla\vec{u}$ in the inertial term, even though the viscous term is multiplied by a small coefficient.

4.6. Stokes equations

At low Reynolds numbers a different nondimensionalization of the pressure, based on the viscosity rather than the inertia, is appropriate. Using

$$\vec{u}^* = \frac{\vec{u}}{U}, \quad p^* = \frac{p}{\rho U^2}, \quad \vec{x}^* = \frac{\vec{x}}{L}, \quad t^* = \frac{Ut}{L},$$

in (2.7), and dropping the stars on the dimensionless variables, we get

$$\begin{aligned} &\mathrm{R}\,(\vec{u}_t + \vec{u}\cdot\nabla\vec{u}) + \nabla p = \Delta\vec{u}, \\ &\nabla\cdot\vec{u} = 0. \end{aligned}$$

Setting $\mathrm{R} = 0$ in these equations, we get the *Stokes equations*,

$$\nabla p = \Delta\vec{u}, \qquad \nabla\cdot\vec{u} = 0. \tag{2.9}$$

These equations provide a good approximation for low Reynolds number flows (although nonuniformities arise in using them on unbounded domains). They are much simpler to analyze than the full Navier-Stokes equations because they are linear.

5. Stokes formula for the drag on a sphere

As an example of the solution of the Stokes equations for low Reynolds number flows, we will derive Stokes' formula (1851) for the drag on a sphere moving at constant velocity through a highly viscous fluid.

It is convenient to retain dimensional variables, so we consider Stokes equations (2.9) in dimensional form

$$\mu\Delta\vec{u} = \nabla p, \qquad \nabla\cdot\vec{u} = 0. \tag{2.10}$$

We note for future use that we can eliminate the pressure from (2.10) by taking the curl of the momentum equation, which gives

$$\Delta \operatorname{curl}\vec{u} = 0. \tag{2.11}$$

Before considering axisymmetric Stokes flow past a sphere, it is useful to look at the two-dimensional equations. Using Cartesian coordinates with $\vec{x} = (x, y)$ and $\vec{u} = (u, v)$, we may write (2.10) as 3×3 system for (u, v, p):

$$\mu\Delta u = p_x, \qquad \mu\Delta v = p_y, \qquad u_x + v_y = 0.$$

Here, $\Delta = \partial_x^2 + \partial_y^2$ is the two-dimensional Laplacian. In a simply connected region, the incompressibility condition implies that we may introduce a streamfunction $\psi(x, y)$ such that $u = \psi_y$ and $v = -\psi_x$. The momentum equation then becomes

$$\mu\Delta\psi_y = p_x, \qquad \Delta\psi_x = -p_y.$$

The elimination of p by cross-differentiation implies that ψ satisfies the biharmonic equation

$$\Delta^2\psi = 0.$$

Thus, the two-dimensional Stokes equations reduce to the biharmonic equation.

Similar considerations apply to axisymmetric flows, although the details are more complicated. We will therefore give a direct derivation of the solution for flow past a sphere, following Landau and Lifshitz [**34**].

We denote the radius of the sphere by a, and adopt a reference frame moving with the sphere. In this reference frame, the sphere is at rest and the fluid velocity far away from the sphere approaches a constant velocity $\vec{U}$. The pressure also approaches a constant, which we may take to be zero without loss of generality.

The appropriate boundary condition for the flow of a viscous fluid past a solid, impermeable boundary is the no-slip condition that the velocity of the fluid is equal to the velocity of the body. Roughly speaking, this means that a viscous fluid 'sticks' to a solid boundary.

Let $\vec{x}$ denote the position vector from the center of the sphere and $r = |\vec{x}|$ the distance. We want to solve the Stokes equations (2.10) for $\vec{u}(\vec{x})$, $p(\vec{x})$ in the exterior of the sphere $a < r < \infty$, subject to the no-slip condition on the sphere,

$$\vec{u}(\vec{x}) = 0 \qquad \text{at } r = a, \tag{2.12}$$

and the uniform-flow condition at infinity,

$$\vec{u}(\vec{x}) \sim \vec{U}, \quad p(\vec{x}) \to 0 \qquad \text{as } r \to \infty. \tag{2.13}$$

First, we will solve for the velocity. Since $\vec{u}$ is divergence free and the exterior of a sphere is simply connected, we can write it as

$$\vec{u}\,(\vec{x}) = \vec{U} + \operatorname{curl}\vec{A}\,(\vec{x}) \tag{2.14}$$

where $\vec{A}$ is a vector-streamfunction for the deviation of the flow from the uniform flow. We can always choose $\vec{A}$ to be divergence free, and we require that the derivatives of $\vec{A}$ approach zero at infinity so that $\vec{u}$ approaches $\vec{U}$.

We will show that we can obtain the solution by choosing $\vec{A}$ to be of the form

$$\vec{A}(\vec{x}) = \nabla f(r) \times \vec{U} \tag{2.15}$$

for a suitable scalar valued function $f(r)$ which we will determine. This form for $\vec{A}$ is dictated by linearity and symmetry considerations: since the Stokes equations are linear, the solution must be linear in the velocity $\vec{U}$; and the solution must be invariant under rotations about the axis parallel to $\vec{U}$ through the center of the sphere, and under rotations of $\vec{U}$.

Using the vector identity

$$\operatorname{curl}\left(f\vec{F}\right) = \nabla f \times \vec{F} + f \operatorname{curl}\vec{F},$$

and the fact that $\vec{U}$ is constant, we may also write (2.15) as

$$\vec{A} = \operatorname{curl}\left(f\vec{U}\right), \tag{2.16}$$

which shows, in particular, that $\nabla \cdot \vec{A} = 0$.

By use of the vector identity

$$\operatorname{curl}\operatorname{curl}\vec{F} = \nabla\left(\nabla \cdot \vec{F}\right) - \Delta\vec{F}, \tag{2.17}$$

and (2.15), we find that

$$\operatorname{curl}\vec{u} = \operatorname{curl}\operatorname{curl}\vec{A} = -\Delta\vec{A} = -\Delta\left(\nabla f \times \vec{U}\right).$$

Using this result in (2.11), we find that

$$\Delta^2\left(\nabla f \times \vec{U}\right) = 0.$$

Since $\vec{U}$ is constant, it follows that

$$\nabla\left(\Delta^2 f\right) \times \vec{U} = 0.$$

Since f depends only on r, this equation implies that $\nabla\left(\Delta^2 f\right) = 0$, so $\Delta^2 f$ is constant. Since the derivatives of $\vec{A}$ decay at infinity, $\Delta^2 f$ must also decay, so the constant is zero, and therefore f satisfies the biharmonic equation

$$\Delta^2 f = 0. \tag{2.18}$$

Writing $g = \Delta f$, which is a function of $r = |\vec{x}|$, and using the expression for the three-dimensional Laplacian in spherical-polar coordinates, we get

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dg}{dr}\right) = 0.$$

Integrating this equation, we get $g(r) = 2b/r + c$ where b, c are constant of integration. Since $\Delta f \to 0$ as $r \to \infty$, we must have $c = 0$, so

$$\Delta f = \frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{df}{dr}\right) = \frac{2b}{r}.$$

Integrating this equation and neglecting an additive constant, which involves no loss of generality because $\vec{A}$ depends only on ∇f, we get

$$f(r) = br + \frac{c}{r} \tag{2.19}$$

where c is another constant of integration.

Using this expression for f in (2.15), then using the result in (2.14), we find that

$$\vec{u}\,(\vec{x}) = \vec{U} - \frac{b}{r}\left[\vec{U} + \frac{1}{r^2}\left(\vec{U}\cdot\vec{x}\right)\vec{x}\right] + \frac{c}{r^3}\left[\frac{3}{r^2}\left(\vec{U}\cdot\vec{x}\right)\vec{x} - \vec{U}\right]. \tag{2.20}$$

This velocity field satisfies the boundary condition at infinity (2.13). Imposing the boundary condition (2.12) on the sphere, we get

$$\left(1 - \frac{b}{a} - \frac{c}{a^3}\right)\vec{U} + \frac{1}{a^3}\left(\frac{3c}{a^2} - b\right)\left(\vec{U}\cdot\vec{x}\right)\vec{x} = 0 \qquad \text{when } |\vec{x}| = a.$$

This condition is satisfied only if the coefficient of each term vanishes, which gives

$$b = \frac{3a}{4}, \qquad c = \frac{a^3}{4}.$$

Thus, from (2.19), the solution for f is

$$f(r) = \frac{3ar}{4}\left(1 + \frac{a^2}{3r^2}\right), \tag{2.21}$$

and, from (2.20), the solution for the velocity field is

$$\vec{u}\,(\vec{x}) = \left(1 - \frac{3a}{4r} - \frac{a^3}{4r^3}\right)\vec{U} + \frac{1}{r^2}\left(\frac{3a^3}{4r^3} - \frac{3a}{4r}\right)\left(\vec{U}\cdot\vec{x}\right)\vec{x}. \tag{2.22}$$

Remark 2.10. A noteworthy feature of this solution is its slow decay to the uniform flow. The difference

$$\vec{U} - \vec{u}\,(\vec{x}) \sim \frac{3a}{4r}\left[\vec{U} + \frac{1}{r^2}\left(\vec{U}\cdot\vec{x}\right)\vec{x}\right]$$

is of the order $1/r$ as $r \to \infty$. This makes the analysis of nondilute suspensions of particles difficult, even with linearity of the Stokes equations, because the hydrodynamic interactions between particles have a long range.

To get the pressure, we first compute $\Delta\vec{u}$. Using (2.16) in (2.14), and applying the vector identity (2.17), we get

$$\begin{aligned}\vec{u} &= \vec{U} + \operatorname{curl}\operatorname{curl}\left(f\vec{U}\right)\\ &= \vec{U} + \nabla\left[\nabla\cdot\left(f\vec{U}\right)\right] - (\Delta f)\,\vec{U}.\end{aligned}$$

Taking the Laplacian of this equation, then using the identity $\nabla\cdot\left(f\vec{U}\right) = \vec{U}\cdot\nabla f$ and the fact that f satisfies the biharmonic equation (2.18), we get

$$\Delta\vec{u} = \nabla\Delta\left(\vec{U}\cdot\nabla f\right).$$

Use of this expression in the momentum equation in (2.10) gives

$$\nabla\left[\mu\Delta\left(\vec{U}\cdot\nabla f\right) - p\right] = 0.$$

It follows that the expression inside the gradient is constant, and from (2.13) the constant is zero. Therefore,

$$p = \mu\Delta\left(\vec{U}\cdot\nabla f\right).$$

Using (2.21) in this equation, we find the explicit expression

$$p = -\left(\frac{3\mu a}{2r^3}\right)\vec{U}\cdot\vec{x}. \tag{2.23}$$

Thus, (2.22) and (2.23) is the solution of (2.10) subject to the boundary conditions (2.13)–(2.12).

A primary quantity of interest is the drag force F exerted by the fluid on the sphere. This force is given by integrating the stress over the surface $\partial\Omega$ of the sphere:

$$\vec{F} = \int_{\partial\Omega}\mathbf{T}\vec{n}\,dS.$$

Here, $\vec{n}$ is the unit outward normal to the sphere, and $\mathbf{T}$ is the Cauchy stress tensor, given from (2.4) by

$$\mathbf{T} = -p\mathbf{I} + \mu\left(\nabla\vec{u} + \nabla\vec{u}^{\top}\right).$$

A direct calculation, whose details we omit, shows that the force is in the direction of $\vec{U}$ with magnitude

$$F = 6\pi\mu a U, \tag{2.24}$$

where U is the magnitude of $\vec{U}$.

This expression for the drag on a spherical particle is found to be in excellent agreement with experimental results if $\mathrm{R} < 0.5$, where

$$\mathrm{R} = \frac{2aU}{\nu}$$

is the Reynolds numbers based on the particle diameter, and $\nu = \mu/\rho_0$ is the kinematic viscosity, as before.

For example, consider a particle of radius a and density ρ_p falling under gravity in a fluid of density ρ_0. At the terminal velocity U, the viscous drag must balance the gravitational buoyancy force, so

$$6\pi\mu a U = \frac{4}{3}\pi a^3\left(\rho_p - \rho_0\right)g$$

where g is the acceleration due to gravity. This gives

$$U = \frac{2a^2 g}{9\nu}\left(\frac{\rho_p}{\rho_0} - 1\right)$$

The corresponding Reynolds number is

$$\mathrm{R} = \frac{4a^3 g}{9\nu^2}\left(\frac{\rho_p}{\rho_0} - 1\right).$$

For a water droplet falling through air [**9**], we have $\rho_p/\rho_0 \approx 780$ and $\nu \approx 15\,\mathrm{mm\,s^{-1}}$. This gives a Reynolds number of approximately $1.5\times10^4\,a^3$ where a is measured in mm. Thus, Stokes formula is applicable when $a \le 0.04\,\mathrm{mm}$, corresponding to droplets in a fine mist.

6. Kolmogorov's 1941 theory of turbulence

> Finally, if we are to list the reasons for studying homogeneous turbulence, we should add that it is a profoundly interesting physical phenomenon which still defies satisfactory mathematical analysis; this is, of course, the most compelling reason.[1]

High-Reynolds number flows typically exhibit an extremely complicated behavior called turbulence. In fact, Reynolds first introduced the Reynolds number in connection with his studies on transition to turbulence in pipe flows in 1895. The analysis and understanding of turbulence remains a fundamental challenge. There is, however, no precise definition of fluid turbulence, and there are many different kinds of turbulent flows, so this challenge is likely to be one with many different parts.

In 1941, Kolmogorov proposed a simple dimensional argument that is one of the basic results about turbulence. To explain his argument, we begin by describing an idealized type of turbulence called homogeneous, isotropic turbulence.

6.1. Homogeneous, isotropic turbulence

Following Batchelor [**8**], let us imagine an infinite extent of fluid in turbulent motion. This means, first, that the fluid velocity depends on a large range of length scales; we denote the smallest length scale (the 'dissipation' length scale) by λ_d and the largest length scale (the 'integral' length scale) by L. And, second, that the fluid motion is apparently random and not reproducible in detail from one experiment to the next.

We therefore adopt a probabilistic description, and suppose that a turbulent flow is described by a probability measure on solutions of the Navier-Stokes equations such that expected values of the fluid variables with respect to the measure agree with appropriate averages of the turbulent flow.

This probabilistic description is sometimes interpreted as follows: we have an 'ensemble' of many different fluid flows — obtained, for example, by repeating the same experiment many different times — and each member of the ensemble corresponds to a flow chosen 'at random' with respect to the probability measure.

A turbulent flow is said to be homogeneous if its expected values are invariant under spatial translations — meaning that, on average, it behaves the same way at each point in space — and isotropic if its expected values are also independent of spatial rotations. Similarly, the flow is stationary if its expected values are invariant under translations in time. Of course, any particular realization of the flow varies in space and time.

Homogeneous, isotropic, stationary turbulence is rather unphysical. Turbulence is typically generated at boundaries, and the properties of the flow vary with distance from the boundary or other large-scale features of the flow geometry. Moreover, turbulence dissipates energy at a rate which appears to be nonzero even in the limit of infinite Reynolds number. Thus, some sort of forcing (usually at the integral length scale) that adds energy to the fluid is required to maintain stationary turbulence. Nevertheless, appropriate experimental configurations (for example, high-Reynolds number flow downstream of a metal grid) and numerical configurations (for example, direct numerical simulations on a 'box' with periodic

[1] G. K. Batchelor, *The Theory of Homogeneous Turbulence.*

boundary conditions and a suitable applied force) provide a good approximation to homogeneous, isotropic turbulence.

6.2. Correlation functions and the energy spectrum

We denote expected values by angular brackets $\langle\cdot\rangle$. In a homogeneous flow the two-point correlation

$$Q = \langle \vec{u}\,(\vec{x},t)\cdot\vec{u}\,(\vec{x}+\vec{r},t)\rangle \tag{2.25}$$

is a function of the spatial displacement $\vec{r}$, and independent of $\vec{x}$. In a stationary flow it is independent of t. Furthermore, in an isotropic flow, Q is a function only of the magnitude $r = |\vec{r}|$ of the displacement vector.

Note, however, that even in isotropic flow the general correlation tensor

$$\mathbf{Q}\,(\vec{r}) = \langle \vec{u}\,(\vec{x},t)\otimes\vec{u}\,(\vec{x}+\vec{r},t)\rangle\,,$$

with components $Q_{ij} = \langle u_i u_j\rangle$, depends on the vector $\vec{r}$, not just its magnitude, because a rotation of $\vec{r}$ also induces a rotation of $\vec{u}$.

For isotropic turbulence, one can show [**8**] that the two-point correlation (2.25) has the Fourier representation

$$Q\,(r) = 2\int_0^\infty \frac{\sin kr}{kr}\,E(k)\,dk$$

where $E(k)$ is a nonnegative function of the wavenumber magnitude k.

In particular, it follows that

$$\frac{1}{2}\,\langle \vec{u}\,(\vec{x},t)\cdot\vec{u}\,(\vec{x},t)\rangle = \int_0^\infty E(k)\,dk. \tag{2.26}$$

Thus, $E(k)$ may be interpreted as the mean kinetic energy density of the turbulent flow as a function of the wavenumber $0 \le k < \infty$.

6.3. The five-thirds law

In fully developed turbulence, there is a wide range of length scales $\lambda_d \ll \lambda \ll L$ that are much greater than the dissipation length scale and much less than the integral length scale. This range is called the inertial range. The corresponding wavenumbers are $k = 2\pi/\lambda$, with dimension

$$[k] = \frac{1}{L}.$$

It appears reasonable to assume that the components of a turbulent flow which vary over length scales in the inertial range do not depend on the viscosity ν of the fluid or on the integral length scale and velocity.

Kolmogorov proposed that, in the inertial range, the flow statistics depend only on the mean rate per unit mass ϵ at which the turbulent flow dissipates energy. It would not make any difference if we used instead the mean rate of energy dissipation per unit volume, since we would have to nondimensionalize this by the fluid density, to get the mean energy dissipation rate per unit mass. The dimension of this rate is is

$$[\epsilon] = \frac{ML^2}{T^2}\cdot\frac{1}{T}\cdot\frac{1}{M} = \frac{L^2}{T^3}.$$

From (2.26), the spectral energy density has dimension

$$[E(k)] = \frac{L^3}{T^2}.$$

If the only quantities on which $E(k)$ depends are the energy dissipation rate ϵ and the wavenumber k itself, then, balancing dimensions, we must have

$$E(k) = C\epsilon^{2/3}k^{-5/3}, \tag{2.27}$$

where C is a dimensionless constant, called the Kolmogorov constant.

Thus, Kolmogorov's 1941 (K41) theory predicts that the energy spectrum of a turbulent flow in the inertial range has a power-law decay as a function of wavenumber with exponent $-5/3$; this is the "five-thirds law."

The spectral result (2.27) was, in fact, first stated by Oboukhov (1941). Kolmogorov gave a closely related result for spatial correlations:

$$\left\langle \left| \vec{u}\left(\vec{x}+\vec{r},t\right) - \vec{u}\left(\vec{x},t\right)\right|^2 \right\rangle = C\epsilon^{2/3}r^{2/3}.$$

This equation suggests that the velocity of a turbulent flow has a 'rough' spatial dependence in the inertial range, similar to that of a non-differentiable Hölder-continuous function with exponent $1/3$.

Onsager rediscovered this result in 1945, and in 1949 suggested that turbulent dissipation might be described by solutions of the Euler equation that are not sufficiently smooth to conserve energy [**20**]. The possible relationship of non-smooth, weak solutions of the incompressible Euler equations (which are highly non-unique and can even increase in kinetic energy without some kind of additional admissibility conditions) to turbulent solutions of the Navier-Stokes equations remains unclear.

6.4. The Kolmogorov length scale

The only length scale that can be constructed from the dissipation rate ϵ and the kinematic viscosity ν, called the Kolmogorov length scale, is

$$\eta = \left(\frac{\nu^3}{\epsilon}\right)^{1/4}.$$

The K41 theory implies that the dissipation length scale is of the order η.

If the energy dissipation rate is the same at all length scales, then, neglecting order one factors, we have

$$\epsilon = \frac{U^3}{L}$$

where L, U are the integral length and velocity scales. Denoting by R_L the Reynolds number based on these scales,

$$\mathrm{R}_L = \frac{UL}{\nu},$$

it follows that

$$\frac{L}{\eta} = \mathrm{R}_L^{3/4}.$$

Thus, according to this dimensional analysis, the ratio of the largest (integral) length scale and the smallest (dissipation) length scale grows like $\mathrm{R}_L^{3/4}$ as $\mathrm{R}_L \to \infty$.

In order to resolve the finest length scales of a three-dimensional flow with integral-scale Reynolds number R_L, we therefore need on the order of

$$N_L = \mathrm{R}_L^{9/4}$$

independent degrees of freedom (for example, N_L Fourier coefficients of the velocity components). The rapid growth of N_L with R_L limits the Reynolds numbers that can be attained in direct numerical simulations of turbulent flows.

6.5. Validity of the five-thirds law

Experimental observations, such as those made by Grant, Stewart and Moilliet (1962) in a tidal channel between islands off Vancouver, agree well with the five-thirds law for the energy spectrum, and give $C \approx 1.5$ in (2.27). The results of DNS on periodic 'boxes', using up to 4096^3 grid points, are also in reasonable agreement with this prediction.

Although the energy spectrum predicted by the K41 theory is close to what is observed, there is evidence that it is not exactly correct. This would imply that there is something wrong with its original assumptions.

Kolmogorov and Oboukhov proposed a refinement of Kolmogorov's original theory in 1962. It is, in particular, not clear that the energy dissipation rate ϵ should be assumed constant, since the energy dissipation in a turbulent flow itself varies over multiple length scales in a complicated fashion. This phenomenon, called 'intermittency,' can lead to corrections in the five-thirds law [**23**]. All such turbulence theories, however, depend on some kind of initial assumptions whose validity can only be checked by comparing their predictions with experimental or numerical observations.

6.6. The benefits and drawbacks of dimensional arguments

As the above examples from fluid mechanics illustrate, dimensional arguments can lead to surprisingly powerful results, even without a detailed analysis of the underlying equations. All that is required is a knowledge of the quantities on which the problem being studied depends together with their dimensions. This does mean, however, one has to know the basic laws that govern the problem, and the dimensional constants they involve. Thus, contrary to the way it sometimes appears, dimensional analysis does not give something for nothing; it can only give what is put in from the start.

This fact cuts both ways. Many of the successes of dimensional analysis, such as Kolmogorov's theory of turbulence, are the result of an insight into which dimensional parameters play an crucial role in a problem and which parameters can be ignored. Such insights typical depend upon a great deal of intuition and experience, and they may be difficult to justify or prove.[2]

Conversely, it may happen that some dimensional parameters that appear to be so small they can be neglected have a significant effect, in which case scaling laws derived on the basis of dimensional arguments that ignore them are likely to be incorrect.

7. Self-similarity

If a problem depends on more fundamental units than the number of dimensional parameters, then we must use the independent or dependent variables themselves to nondimensionalize the problem. For example, we did this when we used the wavenumber k to nondimensionalize the K41 energy spectrum $E(k)$ in (2.27). In

[2]As Bridgeman ([**10**], p. 5) puts it in his elegant 1922 book on dimensional analysis (well worth reading today): "The untutored savage in the bushes would probably not be able to apply the methods of dimensional analysis to this problem and obtain results that would satisfy us." Hopefully, whatever knowledge may have been lost since then in the area of dimensional analysis has been offset by some gains in cultural sensitivity.

that case, we obtain self-similar solutions that are invariant under the scaling transformations induced by a change in the system of units. For example, in a time-dependent problem the spatial profile of a solution at one instant of time might be a rescaling of the spatial profile at any other time.

These self-similar solutions are often among the few solutions of nonlinear equations that can be obtained analytically, and they can provide valuable insight into the behavior of general solutions. For example, the long-time asymptotics of solutions, or the behavior of solutions at singularities, may be given by suitable self-similar solutions.

As a first example, we use dimensional arguments to find the Green's function of the heat equation.

7.1. The heat equation

Consider the following IVP for the Green's function of the heat equation in $\mathbb{R}^d$:

$$\begin{aligned} u_t &= \nu \Delta u, \\ u(x,0) &= E\delta(x). \end{aligned}$$

Here δ is the delta-function, representing a unit point source at the origin. Formally, we have

$$\int_{\mathbb{R}^d} \delta(x)\, dx = 1, \qquad \delta(x) = 0 \quad \text{for } x \neq 0.$$

The dimensioned parameters in this problem are the diffusivity ν and the energy E of the point source. The only length and times scales are those that come from the independent variables (x,t), so the solution is self-similar.

We have $[u] = \theta$, where θ denotes a unit of temperature. Furthermore, since

$$\int_{\mathbb{R}^d} u(x,0)\, dx = E,$$

we have $[E] = \theta L^d$. The rotational invariance of the Laplacian, and the uniqueness of the solution, implies that the solution must be spherically symmetric. Dimensional analysis then gives

$$u(x,t) = \frac{E}{(\nu t)^{d/2}} f\left(\frac{|x|}{\sqrt{\nu t}}\right).$$

Using this expression for $u(x,t)$ in the PDE, we get an ODE for $f(\xi)$,

$$f'' + \left(\frac{\xi}{2} + \frac{d-1}{\xi}\right) f' + \frac{d}{2} f = 0.$$

We can rewrite this equation as a first-order ODE for $f' + \frac{\xi}{2} f$,

$$\left(f' + \frac{\xi}{2} f\right)' + \frac{d-1}{\xi}\left(f' + \frac{\xi}{2} f\right) = 0.$$

Solving this equation, we get

$$f' + \frac{\xi}{2} f = \frac{b}{\xi^{d-1}},$$

where b is a constant of integration. Solving for f, we get

$$f(\xi) = a e^{-\xi^2/4} + b e^{-\xi^2/4} \int \frac{e^{-\xi^2}}{\xi^{d-1}}\, d\xi,$$

where a s another constant of integration. In order for f to be integrable, we must set $b = 0$. Then

$$u(x,t) = \frac{aE}{(\nu t)^{d/2}} \exp\left(-\frac{|x|^2}{4\nu t}\right).$$

Imposing the requirement that

$$\int_{\mathbb{R}^d} u(x,t)\,dx = E,$$

and using the standard integral

$$\int_{\mathbb{R}^d} \exp\left(-\frac{|x|^2}{2c}\right)\,dx = (2\pi c)^{d/2},$$

we find that $a = (4\pi)^{-d/2}$, and

$$u(x,t) = \frac{E}{(4\pi\nu t)^{d/2}} \exp\left(-\frac{|x|^2}{4\nu t}\right).$$

8. The porous medium equation

In this section, we will further illustrate the use of self-similar solutions by describing a problem for point-source solutions of the porous medium equation, taken from Barenblatt [**7**]. This solution is a one-dimensional version of the radially symmetric self-similar solution of the porous medium equation

$$u_t = \nabla \cdot (u \nabla u)$$

found by Zeldovich and Kompaneets (1950) and Barenblatt (1952).

We consider the flow under gravity of an incompressible fluid in a porous medium, such as groundwater in a rock formation. We suppose that the porous medium sits above a horizontal impermeable stratum, and, to simplify the discussion, that the flow is two-dimensional. It is straightforward to treat three-dimensional flows in a similar way.

Let x and z denote horizontal and vertical spatial coordinates, respectively, where the impermeable stratum is located at $z = 0$. Suppose that the porous medium is saturated with fluid for $0 \le z \le h(x,t)$ and dry for $z > h(x,t)$. If the wetting front $z = h(x,t)$ has small slope, we may use a quasi-one dimensional approximation in which we neglect the z-velocity components of the fluid and average x-velocity components with respect to z.

The volume of fluid (per unit length in the transverse y-direction) in $a \le x \le b$ is given by

$$\int_a^b nh(x,t)\,dx$$

where n is the porosity of the medium. That is, n is the ratio of the open volume in the medium that can be occupied by fluid to the total volume. Typical values of n are 0.3–0.7 for clay, and 0.01, or less, for dense crystalline rocks. We will assume that n is constant, in which case it will cancel from the equations.

Let $u(x,t)$ denote the depth-averaged x-component of the fluid velocity. Conservation of volume for an incompressible fluid implies that for any x-interval $[a,b]$

$$\frac{d}{dt}\int_a^b nh(x,t)\,dx = -\left[nhu\right]_a^b.$$

In differential form, we get

$$h_t = -\left(hu\right)_x \tag{2.28}$$

For slow flows, we can assume that the pressure p in the fluid is equal to the hydrostatic pressure

$$p = \rho_0 g\left(h - z\right).$$

It follows that the total pressure 'head', defined by

$$\frac{p}{\rho_0 g} + z,$$

is independent of z and equal to $h(x,t)$.

According to Darcy's law, the volume-flux (or velocity) of a fluid in a porous medium is proportional to the gradient of the pressure head, meaning that

$$u = -kh_x, \tag{2.29}$$

where k is the permeability, or hydraulic conductivity, of the porous medium.

Remark 2.11. Henri Darcy was a French water works engineer. He published his law in 1856 after conducting experiments on the flow of water through columns of sand, which he carried out in the course of investigating fountains in Dijon.

The permeability k in (2.29) has the dimension of $L^2/(HT)$, where H is the dimension of the head h. Since we measure h in terms of vertical height, k has the dimension of velocity. Typical values of k for consolidated rocks range from 10^{-9} m/day for unfractured metamorphic rocks, to 10^3 m/day for karstic limestone.

Using (2.29) in (2.28), we find that $h(x,t)$ satisfies the *porous medium equation*

$$h_t = k\left(hh_x\right)_x. \tag{2.30}$$

We may alternatively write (2.30) as

$$h_t = \frac{1}{2}k\left(h^2\right)_{xx}.$$

This equation was first considered by Boussinesq (1904).

The porous medium equation is an example of a degenerate diffusion equation. It has a nonlinear diffusivity equal to kh which vanishes when $h = 0$. As we will see, this has the interesting consequence that (2.30) has solutions (corresponding to wetting fronts) that propagate into a region with $h = 0$ at finite speed — behavior one would expect of a wave equation, but not at first sight of a diffusion equation.

8.1. A point source solution

Consider a solution $h(x,t)$ of the porous medium equation (2.30) that approaches an initial point source:

$$h(x,t) \to I\delta(x), \qquad t \to 0^+,$$

where $\delta(x)$ denotes the Dirac delta 'function.' Explicitly, this means that we require

$$\begin{aligned} &h(x,t) \to 0 \qquad \text{as } t \to 0^+ \text{ if } x \neq 0, \\ &\lim_{t\to 0^+} \int_{-\infty}^{\infty} h(x,t)\,dx = I. \end{aligned} \tag{2.31}$$

The delta-function is a distribution, rather than a function. We will not discuss distribution theory here (see [**44**] for an introduction and [**27**] for a detailed account). Instead, we will define the delta-function formally as a 'function' δ with the properties that

$$\delta(x) = 0 \qquad \text{for } x \neq 0,$$
$$\int_{-\infty}^{\infty} f(x)\delta(x)\,dx = f(0)$$

for any continuous function f.

The solution of the porous medium with the initial data (2.31) describes the development of of wetting front due to an instantaneous 'flooding' at the origin by a volume of water I. It provides the long time asymptotic behavior of solutions of the porous medium equation with a concentrated non-point source initial condition $h(x,t) = h_0(x)$ where h_0 is a compactly supported function with integral I.

The dimensional parameters at our disposal in solving this problem are k and I. A fundamental system of units is L, T, H where H is a unit for the pressure head. Since we measure the pressure head h in units of length, it is reasonable to ask why we should use different units for h and x. The explanation is that the units of vertical length used to measure the head play a different role in the model than the units used to measure horizontal lengths, and we should be able to rescale x and z independently.

Equating the dimension of different terms in (2.30), we find that

$$[k] = \frac{L^2}{HT}, \qquad [I] = LH.$$

Since we assume that the initial data is a point source, which does not define a length scale, there are no other parameters in the problem.

Two parameters k, I are not sufficient to nondimensionalize a problem with three fundamental units. Thus, we must also use one of the variables to do so. Using t, we get

$$\left[(kIt)^{1/3}\right] = L, \qquad [t] = T, \qquad \left[\frac{I^{2/3}}{(kt)^{1/3}}\right] = H$$

Dimensional analysis then implies that

$$h(x,t) = \frac{I^{2/3}}{(kt)^{1/3}} F\left(\frac{x}{(kIt)^{1/3}}\right)$$

where $F(\xi)$ is a dimensionless function.

Using this similarity form in (2.30), we find that $F(\xi)$ satisfies the ODE

$$-\frac{1}{3}\left(\xi F' + F\right) = \left(FF'\right)'. \tag{2.32}$$

Furthermore, (2.31) implies that

$$F(\xi) \to 0 \qquad \text{as } |\xi| \to \infty, \tag{2.33}$$

$$\int_{-\infty}^{\infty} F(\xi)\,d\xi = 1. \tag{2.34}$$

Integrating (2.32), we get

$$-\frac{1}{3}\xi F + C = FF' \tag{2.35}$$

where C is a constant of integration.

The condition (2.33) implies that $C = 0$. It then follows from (2.35) that either $F = 0$, or

$$F' = -\frac{1}{3}\xi,$$

which implies that

$$F(\xi) = \frac{1}{6}\left(a^2 - \xi^2\right)$$

where a is a constant of integration.

In order to get a solution that is continuous and approaches zero as $|\xi| \to \infty$, we choose

$$F(\xi) = \begin{cases} \left(a^2 - \xi^2\right)/6 & \text{if } |\xi| < a, \\ 0 & \text{if } |\xi| \geq a. \end{cases}$$

The condition (2.34) then implies that

$$a = \left(\frac{9}{2}\right)^{1/3}.$$

Thus, the solution of (2.30)–(2.31) is given by

$$h(x,t) = \frac{I^{2/3}}{6\left(kt\right)^{1/3}}\left[\left(\frac{9}{2}\right)^{2/3} - \frac{x^2}{\left(kIt\right)^{2/3}}\right] \qquad \text{if } |x| < (9kIt/2)^{1/3}$$

with $h(x,t) = 0$ otherwise.

This solution represents a saturated region of finite width which spreads out at finite speed. The solution is not a classical solution of (2.30) since its derivative h_x has a jump discontinuity at $x = \pm(9kIt/2)^{1/3}$. It can be understood as a weak solution in an appropriate sense, but we will not discuss the details here.

The fact that the solution has length scale proportional to $t^{1/3}$ after time t could have been predicted in advance by dimensional analysis, since $L = (kIt)^{1/3}$ is the only horizontal length scale in the problem. The numerical factors, and the fact that the solution has compact support, depend upon the detailed analytical properties of the porous medium equation equation; they could not be shewn by dimensional analysis.

8.2. A pedestrian derivation

Let us consider an alternative method for finding the point source solution that does not require dimensional analysis, but is less systematic.

First, we remove the constants in (2.30) and (2.31) by rescaling the variables. Defining

$$u\left(x,\bar{t}\right) = \frac{1}{I}h\left(x,t\right), \qquad \bar{t} = kIt,$$

and dropping the bars on $\bar{t}$, we find that $u(x,t)$ satisfies

$$u_t = \left(uu_x\right)_x. \tag{2.36}$$

The initial condition (2.31) becomes

$$\begin{aligned} &u(x,t) \to 0 \qquad \text{as } t \to 0^+ \text{ if } x \neq 0, \\ &\lim_{t\to 0^+} \int_{-\infty}^{\infty} u(x,t)\,dx = 1. \end{aligned} \tag{2.37}$$

We seek a similarity solution of (2.36)–(2.37) of the form

$$u(x,t) = \frac{1}{t^m} f\left(\frac{x}{t^n}\right) \tag{2.38}$$

for some exponents m, n. In order to obtain such a solution, the PDE for $u(x,t)$ must reduce to an ODE for $f(\xi)$. As we will see, this is the case provided that m, n are chosen appropriately.

Remark 2.12. Equation (2.38) is a typical form of a self-similar solution that is invariant under scaling transformations, whether or not they are derived from a change in units. Dimensional analysis of this problem allowed us to deduce that the solution is self-similar. Here, we simply seek a self-similar solution of the form (2.38) and hope that it works.

Defining the similarity variable

$$\xi = \frac{x}{t^n}$$

and using a prime to denote the derivative with respect to ξ, we find that

$$\begin{aligned} u_t &= -\frac{1}{t^{m+1}} \left(mf + n\xi f'\right) \\ (uu_x)_x &= \frac{1}{t^{2m+2n}} \left(ff'\right)'. \end{aligned} \tag{2.39}$$

In order for (2.36) to be consistent with (2.38), the powers of t in (2.39) must agree, which implies that

$$m + 2n = 1. \tag{2.40}$$

In that case, $f(\xi)$ satisfies the ODE

$$\left(ff'\right)' + n\xi f' + mf = 0.$$

Thus, equation (2.36) has a one-parameter family of self-similar solutions. The ODE for similarity solutions is easy to integrate when $m = n$, but it is not as simple to solve when $n \neq m$.

To determine the value of m, n for the point source problem, we compute that, for solutions of the form (2.38),

$$\int_{-\infty}^{\infty} u(x,t)\,dx = t^{n-m} \int_{-\infty}^{\infty} f\left(\xi\right)\,d\xi.$$

Thus, to get a nonzero, finite limit as $t \to 0^+$ in (2.37), we must take $m = n$, and then (2.40) implies that $m = n = 1/3$. We therefore recover the same solution as before.

8.3. Scaling invariance

Let us consider the scaling invariances of the porous medium equation (2.36) in more detail.

We consider a rescaling of the independent and dependent variables given by

$$\tilde{x} = \alpha x, \qquad \tilde{t} = \beta t, \qquad \tilde{u} = \mu u \tag{2.41}$$

where α, β, μ are positive constants. Writing u in terms of $\tilde{u}$ in (2.36) and using the transformation of derivatives

$$\partial_x = \alpha \partial_{\tilde{x}}, \qquad \partial_t = \beta \partial_{\tilde{t}},$$

we find that $\tilde{u}\left(\tilde{x}, \tilde{t}\right)$ satisfies the PDE

$$\tilde{u}_{\tilde{t}} = \frac{\alpha^2}{\beta\mu} \left(\tilde{u}\tilde{u}_{\tilde{x}}\right)_{\tilde{x}}.$$

Thus, the rescaling (2.41) leaves (2.36) invariant if $\alpha^2 = \beta\mu$.

To reformulate this invariance in a more geometric way, let $E = \mathbb{R}^2 \times \mathbb{R}$ be the space with coordinates (x, t, u). For $\alpha, \beta > 0$ define the transformation

$$g\left(\alpha, \beta\right) : E \to E, \qquad g(\alpha, \beta) : (x, t, u) \mapsto \left(\alpha x, \beta t, \frac{\alpha^2}{\beta} u\right). \tag{2.42}$$

Then

$$G = \{g(\alpha, \beta) : \alpha, \beta > 0\} \tag{2.43}$$

forms a two-dimensional Lie group of transformations of E:

$$g(1,1) = I, \qquad g^{-1}\left(\alpha, \beta\right) = g\left(\frac{1}{\alpha}, \frac{1}{\beta}\right),$$
$$g\left(\alpha_1, \beta_1\right) g\left(\alpha_2, \beta_2\right) = g\left(\alpha_1\alpha_2, \beta_1\beta_2\right)$$

where I denotes the identity transformation.

The group G is commutative (in general, Lie groups and symmetry groups are not commutative) and is generated by the transformations

$$(x, t, u) \mapsto \left(\alpha x, t, \alpha^2 u\right), \qquad (x, t, u) \mapsto \left(x, \beta t, \frac{1}{\beta} u\right). \tag{2.44}$$

Abusing notation, we use the same symbol to denote the coordinate u and the function $u(x, t)$. Then the action of $g(\alpha, \beta)$ in (2.42) on $u(x, t)$ is given by

$$u(x, t) \mapsto \frac{\alpha^2}{\beta} u\left(\frac{x}{\alpha}, \frac{t}{\beta}\right). \tag{2.45}$$

This map transforms solutions of (2.36) into solutions. Thus, the group G is a symmetry group of (2.36), which consist of the symmetries that arise from dimensional analysis and the invariance of (2.36) under rescalings of its units.

8.4. Similarity solutions

In general, a solution of an equation is mapped to a different solution by elements of a symmetry group. A *similarity solution* is a solution that is mapped to itself by a nontrivial subgroup of symmetries. In other words, it is a fixed point of the subgroup. Let us consider the case of similarity solutions of one-parameter subgroups of scaling transformations for the porous medium equation; we will show that these are the self-similar solutions considered above.

The one-parameter subgroups of G in (2.43) are given by

$$H_n = \{g\,(\beta^n, \beta) : \beta > 0\} \qquad \text{for } -\infty < n < \infty,$$

and $\{g\,(\alpha, 1) : \alpha > 0\}$. From (2.45), a function $u(x,t)$ is invariant under H_n if

$$u(x,t) = \beta^{2n-1} u\left(\frac{x}{\beta^n}, \frac{t}{\beta}\right)$$

for every $\beta > 0$. Choosing $\beta = t$, we conclude that $u(x,t)$ has the form (2.38) with $m = 1 - 2n$. Thus, we recover the self-similar solutions considered previously.

8.5. Translational invariance

A transformation of the space E of dependent and independent variables into itself is called a *point transformation*. The group G in (2.43) does not include all the point transformations that leave (2.36) invariant. In addition to the scaling transformations (2.44), the space-time translations

$$(x,t,u) \mapsto (x - \delta, t, u)\,, \qquad (x,t,u) \mapsto (x, t - \varepsilon, u)\,, \tag{2.46}$$

where $-\infty < \delta, \varepsilon < \infty$, also leave (2.36) invariant, because the terms in the equation do not depend explicitly on (x,t).

As we will show in Section 9.8, the transformations (2.44) and (2.46) generate the full group of point symmetries of (2.36). Thus, the porous medium equation does not have any point symmetries beyond the obvious scaling and translational invariances. This is not always the case, however. Many equations have point symmetries that would be difficult to find without using the theory of Lie algebras.

Remark 2.13. The one-dimensional subgroups of the two-dimensional group of space-time translations are given by

$$(x,t,u) \mapsto (x - c\varepsilon, t - \varepsilon, u)\,,$$

where c is a fixed constant (and also the space translations $(x,t,u) \mapsto (x - \varepsilon, t, u)$). The similarity solutions that are invariant under this subgroup are the traveling wave solutions

$$u(x,t) = f(x - ct).$$

9. Continuous symmetries of differential equations

Dimensional analysis leads to scaling invariances of a differential equation. As we have seen in the case of the porous medium equation, these invariances form a continuous group, or Lie group, of symmetries of the differential equation.

The theory of Lie groups and Lie algebras provides a systematic method to compute all continuous point symmetries of a given differential equation; in fact, this is why Lie first introduced the the theory of Lie groups and Lie algebras.

Lie groups and algebras arise in many other contexts. In particular, as a result of the advent of quantum mechanics in the early 20^{th}-century, where symmetry considerations are crucial, Lie groups and Lie algebras have become a central part of mathematical physics.

We will begin by describing some basic ideas about Lie groups of transformations and their associated Lie algebras. Then we will describe their application to the computation of symmetry groups of differential equations. See Olver [**40, 41**], whose presentation we follow, for a full account.

9.1. Lie groups and Lie algebras

A manifold of dimension d is a space that is locally diffeomorphic to $\mathbb{R}^d$, although its global topology may be different (think of a sphere, for example). This means that the elements of the manifold may, locally, be smoothly parametrized by d coordinates, say $\left(\varepsilon^1, \varepsilon^2, \ldots, \varepsilon^d\right) \in \mathbb{R}^d$. A Lie group is a space that is both a manifold and a group, such that the group operations (composition and inversion) are smooth functions.

Lie groups almost always arise in applications as transformation groups acting on some space. Here, we are interested in Lie groups of symmetries of a differential equation that act as point transformations on the space whose coordinates are the independent and dependent variables of the differential equation.

The key idea we want to explain first is this: the Lie algebra of a Lie group of transformations is represented by the vector fields whose flows are the elements of the Lie Group. As a result, elements of the Lie algebra are often referred to as 'infinitesimal generators' of elements of the Lie group.

Consider a Lie group G acting on a vector space E. In other words, each $g \in G$ is a map $g : E \to E$. Often, one considers Lie groups of linear maps, which are a subgroup of the general linear group $GL(E)$, but we do not assume linearity here.

Suppose that $E = \mathbb{R}^n$, and write the coordinates of $x \in E$ as $\left(x^1, x^2, \ldots, x^n\right)$. We denote the unit vectors in the coordinate directions by

$$\partial_{x^1}, \quad \partial_{x^2}, \quad \ldots, \quad \partial_{x^n}.$$

That is, we identify vectors with their directional derivatives.

Consider a vector field

$$\vec{v}(x) = \xi^i(x)\partial_{x^i},$$

where we use the summation convention in which we sum over repeated upper and lower indices. The associated flow is a one-parameter group of transformations obtained by solving the system of ODEs

$$\frac{dx^i}{d\varepsilon} = \xi^i\left(x^1, x^2, \ldots, x^n\right) \qquad \text{for } 1 \le i \le n.$$

Explicitly, if $x(\varepsilon)$ is a solution of this ODE, then the flow $g(\varepsilon) : x(0) \mapsto x(\varepsilon)$ maps the initial data at $\varepsilon = 0$ to the solution at 'time' ε.

We denote the flow $g(\varepsilon)$ generated by the vector field $\vec{v}$ by

$$g(\varepsilon) = e^{\varepsilon \vec{v}}.$$

Conversely, given a flow $g(\varepsilon)$, we can recover the vector field that generates it from

$$\vec{v}(x) = \left.\frac{d}{d\varepsilon} g(\varepsilon) \cdot x\right|_{\varepsilon=0}.$$

That is, $\vec{v}(x)$ is the tangent, or velocity, vector of the solution curve through x.

Example 2.14. A linear vector field has the form

$$\vec{v}(x) = a^i_j x^j \partial_{x^i}.$$

Its flow is given by the usual exponential $e^{\varepsilon\vec{v}} = e^{\varepsilon A}$ where the linear transformation A has matrix $\left(a^i_j\right)$.

The flow $e^{\varepsilon\vec{v}}$ of a smooth linear vector field $\vec{v}$ is defined for all $-\infty < \varepsilon < \infty$. The flow of a nonlinear vector field may exists only for sufficiently small values of

ε, which may depend on the initial data. In that case we get a local Lie group of flows. Since we only use local considerations here, we will ignore this complication.

9.2. The Lie bracket

In general, a Lie algebra $\mathfrak{g}$ is a vector space with a bilinear, skew-symmetric bracket operation

$$[\cdot,\cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$$

that satisfies the Jacobi identity

$$[u,[v,w]] + [v,[w,u]] + [w,[u,v]] = 0.$$

The Lie bracket of vector fields $\vec{v}$, $\vec{w}$ is defined by their commutator

$$[\vec{v},\vec{w}] = \vec{v}\vec{w} - \vec{w}\vec{v},$$

where the vector fields are understood as differential operators. Explicitly, if

$$\vec{v} = \xi^i \partial_{x^i}, \qquad \vec{w} = \eta^i \partial_{x^i},$$

then

$$[\vec{v},\vec{w}] = \left(\xi^j \frac{\partial \eta^i}{\partial x^j} - \eta^j \frac{\partial \xi^i}{\partial x^j} \right) \partial_{x^i}.$$

The Lie bracket of vector fields measures the non-commutativity of the corresponding flows:

$$[\vec{v},\vec{w}](x) = \frac{1}{2} \frac{d^2}{d\varepsilon^2} \left(e^{\varepsilon\vec{v}} e^{\varepsilon\vec{w}} e^{-\varepsilon\vec{v}} e^{-\varepsilon\vec{w}} \right) x \Big|_{\varepsilon=0}.$$

One can show that the Lie bracket of any two vector field that generate elements of a Lie group of transformations also generates an element of the Lie group. Thus, the infinitesimal generators of the Lie group form a Lie algebra.

9.3. Transformations of the plane

As simple, but useful, examples of Lie transformation groups and their associated Lie algebras, let us consider some transformations of the plane.

The rotations of the plane $g(\varepsilon) : \mathbb{R}^2 \to \mathbb{R}^2$ are given by

$$g(\varepsilon) : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \cos\varepsilon & -\sin\varepsilon \\ \sin\varepsilon & \cos\varepsilon \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \qquad \text{where } \varepsilon \in \mathbb{T}.$$

These transformations form a representation of the one-dimensional Lie group $SO(2)$ on $\mathbb{R}^2$. They are the flow of the ODE

$$\frac{d}{d\varepsilon} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -y \\ x \end{pmatrix}.$$

The vector field on the right hand side of this equation may be written as

$$\vec{v}(x,y) = -y\partial_x + x\partial_y,$$

and thus

$$g(\varepsilon) = e^{\varepsilon\vec{v}}.$$

The Lie algebra $\mathfrak{so}(2)$ of $SO(2)$ consists of the vector fields

$$\{-\varepsilon y\partial_x + \varepsilon x\partial_y : \varepsilon \in \mathbb{R}\}.$$

The translations of the plane in the direction (a,b)

$$(x,y) \mapsto (x - \varepsilon a, y - \varepsilon b)$$

are generated by the constant vector field

$$a\partial_x + b\partial_y$$

The rotations and translations together form the orientation-preserving Euclidean group of the plane, denoted by $E^+(2)$. The full Euclidean group $E(2)$ is generated by rotations, translations, and reflections.

The Euclidean group is not commutative since translations and rotations do not commute. As a result, the corresponding Lie algebra $\mathfrak{e}(2)$ is not trivial. For example, if $\vec{v} = \partial_x$ is an infinitesimal generator of translations in the x-direction, and $\vec{w} = -y\partial_x + x\partial_y$ is an infinitesimal generator of rotations, then $[\vec{v}, \vec{w}] = \partial_y$ is the infinitesimal generator of translations in the y-direction.

The scaling transformations

$$(x, y) \mapsto (e^{\varepsilon r}x, e^{\varepsilon s}y)$$

are generated by the vector field

$$rx\partial_x + sy\partial_y.$$

Together with the translations and rotations, the scaling transformations generate the conformal group of angle preserving transformations of the plane.

Finally, as a nonlinear example, consider the vector field

$$\vec{v}(x, y) = x^2\partial_x - y^2\partial_y.$$

This generates the local flow

$$(x, y) \mapsto \left(\frac{x}{1 - \varepsilon x}, \frac{y}{1 + \varepsilon y}\right).$$

9.4. Transformations of function

Next, we want to consider the action of point transformations on functions.

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$. We denote the coordinates of the independent variables by $x = \left(x^1, x^2, \ldots, x^n\right) \in \mathbb{R}^n$, and the coordinate of the dependent variable by $u \in \mathbb{R}$. We assume that f is scalar-valued only to simplify the notation; it is straightforward to generalize the discussion to vector-valued functions $f : \mathbb{R}^n \to \mathbb{R}^m$.

Let $E = \mathbb{R}^n \times \mathbb{R}$ be the space with coordinates $\left(x^1, \ldots, x^n, u\right)$. Then the graph Γ_f of f is the subset of E given by

$$\Gamma_f = \{(x, u) \in E : u = f(x)\}.$$

Consider a local one-parameter group of point transformations $g(\varepsilon) : E \to E$ on the space of independent and dependent variables. These transformations induce a local transformation of functions, which we denote in the same way,

$$g(\varepsilon) : f \mapsto g(\varepsilon) \cdot f$$

that maps the graph of f to the graph of $g(\varepsilon) \cdot f$. The global image of the graph of f under $g(\varepsilon)$ need not be a graph; it is, however, locally a graph (that is, in a sufficiently small neighborhood of a point $x \in \mathbb{R}^n$ and for small enough values of ε, when $g(\varepsilon)$ is sufficiently close to the identity).

To express the relationship between f and $\tilde{f} = g \cdot f$ explicitly, we write g as

$$g(\varepsilon) : (x, u) \mapsto (\tilde{x}, \tilde{u}), \qquad \tilde{x} = \tilde{X}(x, u, \varepsilon), \quad \tilde{u} = \tilde{U}(x, u, \varepsilon).$$

Then, since

$$g(\varepsilon) : \{(x, u) : u = f(x)\} \mapsto \{(\tilde{x}, \tilde{u}) : \tilde{u} = \tilde{f}(\tilde{x}, \varepsilon)\},$$

we have

$$\tilde{U}\left(x, f(x), \varepsilon\right) = \tilde{f}\left(\tilde{X}(x, f(x), \varepsilon), \varepsilon\right).$$

This is, in general, a complicated implicit equation for $\tilde{f}$ in terms of f.

Example 2.15. Suppose that $f : \mathbb{R} \to \mathbb{R}$ is the square function $f : x \mapsto x^2$ and $g(\varepsilon) : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \times \mathbb{R}$ is the rotation

$$g(\varepsilon) \cdot (x, u) = ((\cos\varepsilon)x - (\sin\varepsilon)u, (\sin\varepsilon)x + (\cos\varepsilon)u).$$

If $u = f(x)$, then

$$\tilde{x} = (\cos\varepsilon)x - (\sin\varepsilon)x^2, \qquad \tilde{u} = (\sin\varepsilon)x + (\cos\varepsilon)x^2.$$

Solving the first equation for x in terms of $\tilde{x}$, then using the second equation to express $\tilde{u}$ in terms of $\tilde{x}$, we find that $\tilde{u} = \tilde{f}\,(\tilde{x}, \varepsilon)$ where

$$\tilde{f}\,(\tilde{x}, \varepsilon) = \frac{2\tilde{x}^2 \cos\varepsilon + 2\tilde{x}\tan\varepsilon}{1 - 2\tilde{x}\sin\varepsilon + \sqrt{1 - 4\tilde{x}\sin\varepsilon/\cos^2\varepsilon}}$$

Thus, the image of the function $x \mapsto x^2$ under the rotation $g(\varepsilon)$ is the function

$$x \mapsto \frac{2x^2 \cos\varepsilon + 2x\tan\varepsilon}{1 - 2x\sin\varepsilon + \sqrt{1 - 4x\sin\varepsilon/\cos^2\varepsilon}}.$$

Note that this reduces to $x \mapsto x^2$ if $\varepsilon = 0$, and that the transformed function is only defined locally if $\varepsilon \neq 0$.

9.5. Prolongation of transformations

In order to obtain the symmetries of a differential equation, we use a geometric formulation of how the derivatives of a function transform under point transformations.

To do this, we introduce a space $E^{(k)}$, called the k^{th} *jet space*, whose coordinates are the independent variables, the dependent variable, and the derivatives of the dependent variable of order less than or equal to k.

We will use multi-index notation for partial derivatives. A multi-index α is an n-tuple

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$$

where each $\alpha_i = 0, 1, 2, \ldots$ is a nonnegative integer. The α-partial derivative of a function $f : \mathbb{R}^n \to \mathbb{R}$ is

$$\partial^\alpha f = \partial_{x^1}^{\alpha_1}\, \partial_{x^2}^{\alpha_2}\, \ldots\, \partial_{x^n}^{\alpha_n}\, f.$$

This partial derivative has order $|\alpha|$ where

$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n.$$

We define $E^{(k)}$ to be the space with coordinates $(x, u, \partial^\alpha u)$ where α runs over all multi-indices with $1 \leq |\alpha| \leq k$. When convenient, we will use alternative notations for the partial-derivative coordinates, such as u_{x^i} for $\partial_{x^i} u$ and u_α for $\partial^\alpha u$.

Example 2.16. Written out explicitly, the coordinates on the first-order jet space $E^{(1)}$ are $\left(x^1, x^2, \ldots, x^n, u, u_{x^1}, u_{x^2}, \ldots, u_{x^n}\right)$. Thus, $E^{(1)}$ has dimension $(2n + 1)$.

Example 2.17. For functions $u = f(x, y)$ of two independent variables, the second-order jet space $E^{(2)}$ has coordinates $(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy})$.

Point transformations induce a map of functions to functions, and therefore they induce maps of the derivatives of functions and of the jet spaces.

Specifically, suppose that $g(\varepsilon) : E \to E$ is a point transformation. We extend, or prolong $g(\varepsilon)$, to a transformation

$$\mathbf{pr}^{(k)} g(\varepsilon) : E^{(k)} \to E^{(k)}$$

in the following way. Given a point $(x, u, \partial^\alpha u) \in E^{(k)}$, pick a function $f : \mathbb{R}^n \to \mathbb{R}$ whose value at x is u and whose derivatives at x are $\partial^\alpha u$, meaning that

$$f(x) = u, \qquad \partial^\alpha f(x) = \partial^\alpha u \qquad \text{for } 1 \leq |\alpha| \leq k.$$

For example, we could choose f to be a polynomial of degree k.

Suppose that $g(\varepsilon) \cdot (x, u) = (\tilde{x}, \tilde{u})$ is the image of $(x, u) \in E$ under $g(\varepsilon)$ and $\tilde{f} = g(\varepsilon) \cdot f$ is the image of the function f. We define the image of the jet-coordinates by

$$\widetilde{\partial^\alpha u} = \tilde{\partial}^\alpha \tilde{f}\,(\tilde{x})\,.$$

That is, they are the values of the derivatives of the transformed function $\tilde{f}\,(\tilde{x})$. One can show that these values do not depend on a particular choice of the function f, so this gives a well-defined map $\mathbf{pr}^{(k)} g(\varepsilon)$ on $E^{(k)}$ such that

$$\mathbf{pr}^{(k)} g(\varepsilon) : (x, u, \partial^\alpha u) \mapsto \left(\tilde{x}, \tilde{u}, \widetilde{\partial^\alpha u}\right).$$

9.6. Prolongation of vector fields

Suppose that $g(\varepsilon) : E \to E$ is generated by the vector field

$$\vec{v}(x, u) = \xi^i(x, u)\partial_{x^i} + \varphi(x, u)\partial_u. \tag{2.47}$$

Then, writing the coordinates of $E^{(k)}$ as (x, u, u_α), the prolonged transformation

$$\mathbf{pr}^{(k)} g(\varepsilon) : E^{(k)} \to E^{(k)}$$

is generated by a vector field $\mathbf{pr}^{(k)} \vec{v}$ on $E^{(k)}$. This prolonged vector field has the form

$$\mathbf{pr}^{(k)} \vec{v} = \xi^i \partial_{x^i} + \varphi \partial_u + \sum_{|\alpha|=1}^{k} \varphi^\alpha \partial_{u_\alpha},$$

where the φ^α are suitable coefficient functions, which are determined by $\vec{v}$.

The prolongation formula expresses the coefficients φ^α of the prolonged vector field in terms of the coefficients ξ^i, φ of the original vector field. We will state the result here without proof (see [**40**] for a derivation).

To write the prolongation formula in a compact form — see (2.49) below — we define the total derivative $D_{x^i} F : E^{(k)} \to \mathbb{R}$ of a function $F : E^{(k)} \to \mathbb{R}$ with respect to an independent variable x^i by

$$D_{x^i} F = \partial_{x^i} F + \sum_{|\alpha|=0}^{k} u_{\alpha,i} \partial_{u_\alpha} F.$$

Here, we use the notation

$$u_{\alpha,i} = \partial_{x^i} \partial^\alpha u$$

to denote the coordinate of the corresponding derivative. That is, $u_{\alpha,i} = u_\beta$ where $\beta_i = \alpha_i + 1$ and $\beta_j = \alpha_j$ for $j \neq i$.

In other words, the total derivative $D_{x^i}F$ of F with respect to x^i is what we would obtain by differentiating F with respect to x^i *after* the coordinates u, u_α have been evaluated at a function of x and its derivatives.

If $\alpha = (\alpha_1, \dots, \alpha_n)$ is a multi-index, we define the α-total derivative by

$$D^\alpha = D_{x^1}^{\alpha_1} D_{x^2}^{\alpha_2} \dots D_{x^n}^{\alpha_n}.$$

Total derivatives commute, so the order in which we take them does not matter.

Finally, we define the characteristic $Q : E^{(1)} \to \mathbb{R}$ of the vector field (2.47) by

$$Q\left(x, u, \partial u\right) = \varphi(x, u) - \xi^i(x, u) u_{x^i},$$

where the summation convention is understood, and $\partial u = (u_{x^1}, \dots, u_{x^n})$ is the first-derivative coordinate.

Then the k^{th}-prolongation of the vector field (2.47) is given by

$$\mathbf{pr}^{(k)} \vec{v} = \xi^i \partial_{x^i} + \varphi \partial_u + \sum_{|\alpha|=1}^{k} \varphi^\alpha \partial_{u_\alpha}, \tag{2.48}$$

$$\varphi^\alpha = D^\alpha Q + \xi^i u_{\alpha,i}. \tag{2.49}$$

This is the main result needed for the algebraic computation of symmetries of a differential equation. See Olver [**40**] for the prolongation formula for systems.

9.7. Invariance of a differential equation

A k^{th} order differential equation for a real-valued function $u(x)$ may be written as

$$F\left(x, u, \partial^\alpha u\right) = 0$$

where $F : E^{(k)} \to \mathbb{R}$ and $1 \le |\alpha| \le k$. Here, we abuse notation and use the same symbols for the coordinates u, $\partial^\alpha u$ and the functions $u(x)$, $\partial^\alpha u(x)$.

A local point transformation $g(\varepsilon) : E \to E$ is a symmetry of the differential equation if $g(\varepsilon) \cdot u$ is a solution whenever u is a solution. This means that, for all ε in the neighborhood of 0 for which $g(\varepsilon)$ is defined, we have

$$F\left(\mathbf{pr}^{(k)} g(\varepsilon) \cdot (x, u, \partial^\alpha u)\right) = F\left(x, u, \partial^\alpha u\right) \qquad \text{on } F\left(x, u, \partial^\alpha u\right) = 0. \tag{2.50}$$

Suppose that $g(\varepsilon) = e^{\varepsilon \vec{v}}$. Then, differentiating (2.50) with respect to ε and setting $\varepsilon = 0$, we conclude that

$$\mathbf{pr}^{(k)} \vec{v} \cdot F\left(x, u, \partial^\alpha u\right) = 0 \qquad \text{on } F\left(x, u, \partial^\alpha u\right) = 0, \tag{2.51}$$

where $\mathbf{pr}^{(k)} \vec{v}$ acts on F by differentiation. Conversely, if F satisfies the 'maximal rank' condition $DF \neq 0$ on $F = 0$, which rules out degenerate ways of the equation such as $F^2 = 0$, we may integrate the infinitesimal invariance condition (2.51) to obtain (2.50).

The condition (2.51) is called the *determining equation* for the infinitesimal symmetries of the differential equation. It is typically a large, over-determined system of equations for $\xi^i(x, u)$, $\varphi(x, u)$ and their derivatives, which is straightforward (though tedious) to solve.

Thus, in summary, to compute the point symmetries of a differential equation

$$F\left(x, u, \partial^\alpha u\right) = 0$$

we use the prolongation formula (2.48)–(2.49) to write down the infinitesimal invariance condition (2.51), solve the resulting equations for $\xi^i(x, u)$ and $\varphi(x, u)$, then integrate the vector fields (2.47) to obtain the symmetries $g = e^{\vec{v}}$.

9.8. Porous medium equation

Let us return to the porous medium equation (2.36).

The space E of independent and dependent variables has coordinates (x, t, u). We may write the equation as

$$F\left(u, u_x, u_t, u_{xx}\right) = 0 \tag{2.52}$$

where $F : E^{(2)} \to \mathbb{R}$ is given by

$$F\left(u, u_x, u_t, u_{xx}\right) = -u_t + u_x^2 + u u_{xx}.$$

A vector field $\vec{v}$ on E is given by

$$\vec{v}(x,t,u) = \xi(x,t,u)\partial_x + \tau(x,t,u)\partial_t + \varphi(x,t,u)\partial_u. \tag{2.53}$$

From (2.48), the second prolongation of $\vec{v}$ has the form

$$\mathbf{pr}^{(2)}\vec{v} = \xi\partial_x + \tau\partial_t + \varphi\partial_u + \varphi^x\partial_{u_x} + \varphi^t\partial_{u_t} + \varphi^{xx}\partial_{u_{xx}} + \varphi^{xt}\partial_{u_{xt}} + \varphi^{tt}\partial_{u_{tt}}.$$

The infinitesimal invariance condition (2.51) applied to (2.52) gives

$$-\varphi^t + 2u_x\varphi^x + \varphi u_{xx} + u\varphi^{xx} = 0 \qquad \text{on } u_t = u_x^2 + u u_{xx}. \tag{2.54}$$

From (2.49), we have

$$\begin{aligned}
\varphi^x &= D_x Q + \xi u_{xx} + \tau u_{xt},\\
\varphi^t &= D_t Q + \xi u_{xt} + \tau u_{tt},\\
\varphi^{xx} &= D_x^2 Q + \xi u_{xxx} + \tau u_{xxt},\\
\varphi^{xt} &= D_x D_t Q + \xi u_{xxt} + \tau u_{xtt},\\
\varphi^{tt} &= D_t^2 Q + \xi u_{xtt} + \tau u_{ttt},
\end{aligned}$$

where the characteristic Q of (2.53) is given by

$$Q\left(x, t, u, u_t, u_x\right) = \varphi\left(x, t, u\right) - \xi\left(x, t, u\right) u_x - \tau\left(x, t, u\right) u_t,$$

and the total derivatives D_x, D_t of a function $f\left(x, t, u, u_x, u_t\right)$ are given by

$$\begin{aligned}
D_x f &= f_x + u_x f_u + u_{xx} f_{u_x} + u_{xt} f_{u_t},\\
D_t f &= f_t + u_t f_u + u_{xt} f_{u_x} + u_{tt} f_{u_t}.
\end{aligned}$$

Expanding the total derivatives of Q, we find that

$$\begin{aligned}
\varphi^x &= \varphi_x + \left(\varphi_u - \xi_x\right) u_x - \tau_x u_t - \xi_u u_x^2 - \tau_u u_x u_t,\\
\varphi^t &= \varphi_t - \xi_t u_x + \left(\varphi_u - \tau_t\right) u_t - \xi_u u_x u_t - \tau_u u_t^2,\\
\varphi^{xx} &= \varphi_{xx} + \left(2\varphi_{xu} - \xi_{xx}\right) u_x - \tau_{xx} u_t + \left(\varphi_{uu} - 2\xi_{xu}\right) u_x^2\\
&\quad - 2\tau_{xt} u_x u_t - \xi_{uu} u_x^3 - \tau_{uu} u_x^2 u_t + \left(\varphi_u - 2\xi_x\right) u_{xx}\\
&\quad - 2\tau_x u_{xt} - 3\xi_u u_x u_{xx} - \tau_u u_t u_{xx} - 2\tau_u u_x u_{xt}.
\end{aligned}$$

We use these expressions in (2.54), replace u_t by $u u_{xx} + u_x^2$ in the result, and equate coefficients of the terms that involve different combinations of the spatial derivatives of u to zero.

The highest derivative terms are those proportional to u_{xt} and $u_x u_{xt}$. Their coefficients are proportional to τ_x and τ_u, respectively, so we conclude that $\tau_x = 0$ and $\tau_u = 0$, which implies that $\tau = \tau(t)$ depends only on t.

The remaining terms that involve second-order spatial derivatives are a term proportional to $u_x u_{xx}$, with coefficient ξ_u, so $\xi_u = 0$ and $\xi = \xi(x, t)$, and a term

proportional to u_{xx}. Equating the coefficient of the latter term to zero, we find that

$$\varphi = (2\xi_x - \tau_t)\, u. \tag{2.55}$$

Thus, φ is a linear function of u.

The terms that are left are either proportional to u_x^2, u_x, or involve no derivatives of u. Equating to zero the coefficients of these terms to zero, we get

$$\begin{aligned} &\tau_t - 2\xi_x + \varphi_u + u\varphi_{uu} = 0, \\ &\xi_t - u\xi_{xx} + 2\varphi_x + 2u\varphi_{xu} = 0, \\ &\varphi_t - u\varphi_{xx} = 0. \end{aligned}$$

The first equation is satisfied by any φ of the form (2.55). The second equation is satisfied if $\xi_t = 0$ and $\xi_{xx} = 0$, which implies that

$$\xi = \varepsilon_1 + \varepsilon_3 x$$

for arbitrary constants ε_1, ε_3. The third equation holds if $\tau_{tt} = 0$, which implies that

$$\tau = \varepsilon_2 + \varepsilon_4 t$$

for arbitrary constants ε_2, ε_3. Equation (2.55) then gives

$$\varphi = (2\varepsilon_3 - \varepsilon_4)\, u$$

Thus, the general form of an infinitesimal generator of a point symmetry of (2.36) is

$$\vec{v}(x,t,u) = (\varepsilon_1 + \varepsilon_3 x)\, \partial_x + (\varepsilon_2 + \varepsilon_4 t)\, \partial_t + (2\varepsilon_3 - \varepsilon_4)\, u\partial_u.$$

We may write this as

$$\vec{v} = \sum_{i=1}^{4} \varepsilon_i \vec{v}_i$$

where the vector fields $\vec{v}_i$ are given by

$$\vec{v}_1 = \partial_x, \qquad \vec{v}_2 = \partial_t \tag{2.56}$$

$$\vec{v}_3 = x\partial_x + 2u\partial_u \qquad \vec{v}_4 = t\partial_t - u\partial_u \tag{2.57}$$

The vector fields $\vec{v}_1$, $\vec{v}_2$ in (2.56) generate the space and time translations

$$(x,t,u) \mapsto (x + \varepsilon_1, t, u), \qquad (x,t,u) \mapsto (x, t + \varepsilon_2, u),$$

respectively. The vector fields $\vec{v}_3$, $\vec{v}_4$ in (2.57) generate the scaling transformations

$$(x,t,u) \mapsto \left(e^{\varepsilon_3} x, t, e^{2\varepsilon_3} u\right) \qquad (x,t,u) \mapsto \left(x, e^{\varepsilon_4} t, e^{-\varepsilon_4} u\right).$$

These are the same as (2.44) with

$$\alpha = e^{\varepsilon_3}, \qquad \beta = e^{\varepsilon_4}.$$

Thus the full point symmetry group of the porous medium equation is four dimensional, and is generated by space and time translations and the two scaling transformations that arise from dimensional analysis.

This result is, perhaps, a little disappointing, since we did not find any new symmetries, although it is comforting to know that there are no other point symmetries to be found. For other equations, however, we can get symmetries that are not at all obvious.

Example 2.18. Consider the one-dimensional heat equation

$$u_t = u_{xx}. \tag{2.58}$$

The determining equation for infinitesimal symmetries is

$$\varphi^t = \varphi^{xx} \qquad \text{on } u_t = u_{xx}.$$

Solving this equation and integrating the resulting vector fields, we find that the point symmetry group of (2.58) is generated by the following transformations [**40**]:

$$\begin{aligned}
&u(x,t) \mapsto u(x-\alpha,t),\\
&u(x,t) \mapsto u(x,t-\beta),\\
&u(x,t) \mapsto \gamma u(x,t),\\
&u(x,t) \mapsto u(\delta x, \delta^2 t),\\
&u(x,t) \mapsto e^{-\epsilon x+\epsilon^2 t} u(x-2\epsilon t,t),\\
&u(x,t) \mapsto \frac{1}{\sqrt{1+4\eta t}} \exp\left[\frac{-\eta x^2}{1+4\eta t}\right] u\left(\frac{x}{1+4\eta t}, \frac{t}{1+4\eta t}\right),\\
&u(x,t) \mapsto u(x,t)+v(x,t),
\end{aligned}$$

where $(\alpha,\dots,\eta)$ are constants, and $v(x,t)$ is an arbitrary solution of the heat equation. The scaling symmetries involving γ and δ can be deduced by dimensional arguments, but the symmetries involving ϵ and η cannot.

As these examples illustrate, given a differential equation it is, in principle, straightforward (but lengthy) to write out the conditions that a vector field generates a point symmetry of the equation, solve these conditions, and integrate the resulting infinitesimal generators to obtain the Lie group of continuous point symmetries of the equation. There are a number of symbolic algebra packages that will do this automatically.

Finally, we note that point symmetries are not the only kind of symmetry one can consider. It is possible to define 'generalized' (also called 'nonclassical' or 'higher') symmetries on infinite-dimensional jet spaces (see [**40**], for an introduction). These are of particular interest in connection with completely integrable equations, such as the Korteweg-de Vries (KdV) equation, which possess 'hidden' symmetries that are not revealed by their point symmetries

LECTURE 3

The Calculus of Variations

> The variational principles of mechanics are firmly rooted in the soil of that great century of Liberalism which starts with Descartes and ends with the French Revolution and which has witnessed the lives of Leibniz, Spinoza, Goethe, and Johann Sebastian Bach. It is the only period of cosmic thinking in the entire history of Europe since the time of the Greeks.[1]

The calculus of variations studies the extreme and critical points of functions. It has its roots in many areas, from geometry to optimization to mechanics, and it has grown so large that it is difficult to describe with any sort of completeness.

Perhaps the most basic problem in the calculus of variations is this: given a function $f : \mathbb{R}^n \to \mathbb{R}$ that is bounded from below, find a point $\bar{x} \in \mathbb{R}^n$ (if one exists) such that

$$f(\bar{x}) = \inf_{x \in \mathbb{R}^n} f(x).$$

There are two main approaches to this problem. One is the 'direct method,' in which we take a sequence of points such that the sequence of values of f converges to the infimum of f, and then try to showing that the sequence, or a subsequence of it, converges to a minimizer. Typically, this requires some sort of compactness to show that there is a convergent subsequence of minimizers, and some sort of lower semi-continuity of the function to show that the limit is a minimizer.

The other approach is the 'indirect method,' in which we use the fact that any interior point where f is differentiable and attains a minimum is a critical, or stationary, point of f, meaning that the derivative of f is zero. We then examine the critical points of f, together with any boundary points and points where f is not differentiable, for a minimum.

Here, we will focus on the indirect method for functionals, that is, scalar-valued functions of functions. In particular, we will derive differential equations, called the Euler-Lagrange equations, that are satisfied by the critical points of certain functionals, and study some of the associated variational problems.

We will begin by explaining how the calculus of variations provides a formulation of one of the most basic systems in classical mechanics, a point particle moving in a conservative force field. See Arnold [**6**] for an extensive account of classical mechanics.

---

[1] Cornelius Lanczos, *The Variational Principles of Mechanics.*

1. Motion of a particle in a conservative force field

Consider a particle of constant mass m moving in n-space dimensions in a spatially-dependent force field $\vec{F}(\vec{x})$. The force field is said to be conservative if

$$\vec{F}(\vec{x}) = -\nabla V(\vec{x})$$

for a smooth potential function $V : \mathbb{R}^n \to \mathbb{R}$, where ∇ denotes the gradient with respect to $\vec{x}$. Equivalently, the force field is conservative if the work done by $\vec{F}$ on the particle as it moves from $\vec{x}_0$ to $\vec{x}_1$,

$$\int_{\Gamma(\vec{x}_0,\vec{x}_1)} \vec{F} \cdot d\vec{x},$$

is independent of the path $\Gamma(\vec{x}_0, \vec{x}_1)$ between the two endpoints.

Abusing notation, we denote the position of the particle at time $a \le t \le b$ by $\vec{x}(t)$. We refer to a function $\vec{x} : [a, b] \to \mathbb{R}^n$ as a particle trajectory. Then, according to Newton's second law, a trajectory satisfies

$$m\ddot{\vec{x}} = -\nabla V(\vec{x}) \tag{3.1}$$

where a dot denotes the derivative with respect to t.

Taking the scalar product of (3.1) with respect to $\dot{\vec{x}}$, and rewriting the result, we find that

$$\frac{d}{dt}\left\{\frac{1}{2}m\left|\dot{\vec{x}}\right|^2 + V(\vec{x})\right\} = 0.$$

Thus, the total energy of the particle

$$E = T\left(\dot{\vec{x}}\right) + V(\vec{x}),$$

where $V(\vec{x})$ is the potential energy and

$$T(\vec{v}) = \frac{1}{2}m|\vec{v}|^2$$

is the kinetic energy, is constant in time.

Example 3.1. The position $x(t) : [a, b] \to \mathbb{R}$ of a one-dimensional oscillator moving in a potential $V : \mathbb{R} \to \mathbb{R}$ satisfies the ODE

$$m\ddot{x} + V'(x) = 0$$

where the prime denotes a derivative with respect to x. The solutions lie on the curves in the $(x, \dot{x})$-phase plane given by

$$\frac{1}{2}m\dot{x}^2 + V(x) = E.$$

The equilibrium solutions are the critical points of the potential V. Local minima of V correspond to stable equilibria, while other critical points correspond to unstable equilibria. For example, the quadratic potential $V(x) = \frac{1}{2}kx^2$ gives the linear simple harmonic oscillator, $\ddot{x} + \omega^2 x = 0$, with frequency $\omega = \sqrt{k/m}$. Its solution curves in the phase plane are ellipses, and the origin is a stable equilibrium.

Example 3.2. The position $\vec{x} : [a, b] \to \mathbb{R}^3$ of a mass m moving in three space dimensions that is acted on by an inverse-square gravitational force of a fixed mass M at the origin satisfies

$$\ddot{\vec{x}} = -GM\frac{\vec{x}}{|\vec{x}|^3},$$

where G is the gravitational constant. The solutions are conic sections with the origin as a focus, as one can show by writing the equations in terms of polar coordinates in the plane of the particle motion motion, and integrating the resulting ODEs.

Example 3.3. Consider n particles of mass m_i and positions $\vec{x}_i(t)$, where $i = 1, 2, ..., n$, that interact in three space dimensions through an inverse-square gravitational force. The equations of motion,

$$\ddot{\vec{x}}_i = -G \sum_{j=1}^{n} m_j \frac{\vec{x}_i - \vec{x_j}}{|\vec{x}_i - \vec{x_j}|^3} \qquad \text{for } 1 \leq i \leq n,$$

are a system of $3n$ nonlinear, second-order ODEs. The system is completely integrable for $n = 2$, when it can be reduced to the Kepler problem, but it is nonintegrable for $n \geq 3$, and extremely difficult to analyze. One of the main results is KAM theory, named after Kolmogorov, Arnold and Moser, on the persistence of invariant tori for nonintegrable perturbations of integrable systems [**6**].

Example 3.4. The configuration of a particle may be described by a point in some other manifold than $\mathbb{R}^n$. For example, consider a pendulum of length ℓ and mass m in a gravitational field with acceleration g. We may describe its configuration by an angle $\theta \in \mathbb{T}$ where $\mathbb{T} = \mathbb{R}/(2\pi\mathbb{Z})$ is the one-dimensional torus (or, equivalently, the circle $\mathbb{S}^1$). The corresponding equation of motion is the pendulum equation

$$\ell\ddot{\theta} + g \sin\theta = 0.$$

1.1. The principle of stationary action

To give a variational formulation of (3.1), we define a function

$$L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R},$$

called the *Lagrangian*, by

$$L(\vec{x}, \vec{v}) = T(\vec{v}) - V(\vec{x}). \tag{3.2}$$

Thus, $L(\vec{x}, \vec{v})$ is the *difference* between the kinetic and potential energies of the particle, expressed as a function of position $\vec{x}$ and velocity $\vec{v}$.

If $\vec{x} : [a, b] \to \mathbb{R}^n$ is a trajectory, we define the *action* of $\vec{x}(t)$ on $[a, b]$ by

$$\mathcal{S}(\vec{x}) = \int_a^b L\left(\vec{x}(t), \dot{\vec{x}}(t)\right) dt. \tag{3.3}$$

Thus, the action $\mathcal{S}$ is a real-valued function defined on a space of trajectories $\{\vec{x} : [a, b] \to \mathbb{R}^n\}$. A scalar-valued function of functions, such as the action, is often called a functional.

The *principle of stationary action* (also called *Hamilton's principle* or, somewhat incorrectly, the *principle of least action*) states that, for fixed initial and final positions $\vec{x}(a)$ and $\vec{x}(b)$, the trajectory of the particle $\vec{x}(t)$ is a stationary point of the action.

To explain what this means in more detail, suppose that $\vec{h} : [a, b] \to \mathbb{R}^n$ is a trajectory with $\vec{h}(a) = \vec{h}(b) = 0$. The directional (or Gâteaux) derivative of $\mathcal{S}$ at $\vec{x}(t)$ in the direction $\vec{h}(t)$ is defined by

$$d\mathcal{S}(\vec{x})\,\vec{h} = \frac{d}{d\varepsilon}\mathcal{S}\left(\vec{x} + \varepsilon\vec{h}\right)\Big|_{\varepsilon=0}. \tag{3.4}$$

The (Fréchet) derivative of $\mathcal{S}$ at $\vec{x}(t)$ is the linear functional $d\mathcal{S}(\vec{x})$ that maps $\vec{h}(t)$ to the directional derivative of $\mathcal{S}$ at $\vec{x}(t)$ in the direction $\vec{h}(t)$.

Remark 3.5. Simple examples show that, even for functions $f : \mathbb{R}^2 \to \mathbb{R}$, the existence of directional derivatives at a point does not guarantee the existence of a Fréchet derivative that provides a local linear approximation of f. In fact, it does not even guarantee the continuity of the function; for example, consider

$$f(x,y) = \frac{xy^2}{x^2+y^4} \qquad \text{if } (x,y) \neq (0,0)$$

with $f(0,0) = 0$. For sufficiently smooth functions, however, such as the action functional we consider here, the existence of directional derivatives does imply the existence of the derivative, and the Gâteaux and Fréchet derivatives agree, so we do not need to worry about the distinction.

A trajectory $\vec{x}(t)$ is a *stationary point* of $\mathcal{S}$ if it is a critical point, meaning that $d\mathcal{S}(\vec{x}) = 0$. Explicitly, this means that

$$\frac{d}{d\varepsilon}\mathcal{S}\left(\vec{x}+\varepsilon\vec{h}\right)\Big|_{\varepsilon=0} = 0$$

for every smooth function $\vec{h} : [a,b] \to \mathbb{R}^n$ that vanishes at $t = a, b$. Thus, small variations in the trajectory of the order ε that keep its endpoints fixed, lead to variations in the action of the order ε^2.

Remark 3.6. Remarkably, the motion of any conservative, classical physical system can be described by a principle of stationary action. Examples include ideal fluid mechanics, elasticity, magnetohydrodynamics, electromagnetics, and general relativity. All that is required to specify the dynamics of a system is an appropriate configuration space to describe its state and a Lagrangian.

Remark 3.7. This meaning of the principle of stationary action is rather mysterious, but we will verify that it leads to Newton's second law. One way to interpret the principle is that it expresses a lack of distinction between different forms of energy (kinetic and potential): any variation of a stationary trajectory leads to an equal gain, or loss, of kinetic and potential energies. An alternative explanation, from quantum mechanics, is that the trajectories with stationary action are the ones with a minimal cancelation of quantum-mechanical amplitudes. Whether this makes the principle less, or more, mysterious is not so clear.

1.2. Equivalence with Newton's second law

To derive the differential equation satisfied by a stationary point $\vec{x}(t)$ of the action $\mathcal{S}$ defined in (3.2)–(3.3), we differentiate the equation

$$\mathcal{S}\left(\vec{x}+\varepsilon\vec{h}\right) = \int_a^b \left\{\frac{1}{2}m\left|\dot{\vec{x}}(t)+\varepsilon\dot{\vec{h}}(t)\right|^2 - V\left(\vec{x}(t)+\varepsilon\vec{h}(t)\right)\right\}\,dt$$

with respect to ε, and set $\varepsilon = 0$, as in (3.4). This gives

$$d\mathcal{S}(\vec{x})\,\vec{h} = \int_a^b \left\{m\dot{\vec{x}}\cdot\dot{\vec{h}} - \nabla V(\vec{x})\cdot\vec{h}\right\}\,dt.$$

Integrating the first term by parts, and using the fact that the boundary terms vanish because $\vec{h}(a) = \vec{h}(b) = 0$, we get

$$d\mathcal{S}(\vec{x})\,\vec{h} = -\int_a^b \left\{ m\ddot{\vec{x}} + \nabla V(\vec{x}) \right\} \cdot \vec{h}\, dt. \tag{3.5}$$

If this integral vanishes for arbitrary $\vec{h}(t)$, it follows from the du Bois-Reymond lemma (1879) that the integrand vanishes. Thus, $\vec{x}(t)$ satisfies

$$m\ddot{\vec{x}} + \nabla V(\vec{x}) = 0$$

for $a \leq t \leq b$. Hence, we recover Newton's second law (3.1).

1.3. The variational derivative

A convenient way to write the derivative of the action is in terms of the variational, or functional, derivative. The variational derivative of $\mathcal{S}$ at $\vec{x}(t)$ is the function

$$\frac{\delta \mathcal{S}}{\delta \vec{x}} : [a,b] \to \mathbb{R}^n$$

such that

$$d\mathcal{S}(\vec{x})\,\vec{h} = \int_a^b \frac{\delta \mathcal{S}}{\delta \vec{x}(t)} \cdot \vec{h}(t)\, dt.$$

Here, we use the notation

$$\frac{\delta \mathcal{S}}{\delta \vec{x}(t)}$$

to denote the value of the variational derivative at t. Note that the variational derivative depends on the trajectory $\vec{x}$ at which we evaluate $d\mathcal{S}(\vec{x})$, although the notation does not show this explicitly.

Thus, for the action functional (3.2)–(3.3), equation (3.5) implies that

$$\frac{\delta \mathcal{S}}{\delta \vec{x}} = -\left\{ m\ddot{\vec{x}} + \nabla V(\vec{x}) \right\}.$$

A trajectory $\vec{x}(t)$ is a stationary point of $\mathcal{S}$ if the variational derivative of $\mathcal{S}$ vanishes at $\vec{x}(t)$.

The variational derivative of a functional is analogous to the gradient of a function. If $f : \mathbb{R}^n \to \mathbb{R}$ is a scalar-valued function on n-dimensional Euclidean space, then the gradient ∇f is defined by

$$\frac{d}{d\varepsilon} f\left(\vec{x} + \varepsilon \vec{h}\right)\Big|_{\varepsilon=0} = \nabla f(\vec{x}) \cdot \vec{h}$$

where '$\cdot$' denotes the Euclidean inner product. Thus, we use the inner product to identify the derivative at a point, which is is a linear map belonging to the dual space of $\mathbb{R}^n$, with a corresponding gradient vector belonging to $\mathbb{R}^n$. For the variational derivative, we replace the Euclidean inner product of vectors by the L^2-inner product of functions,

$$\langle \vec{x}, \vec{y} \rangle = \int_a^b \vec{x}(t) \cdot \vec{y}(t)\, dt,$$

and define the variational derivative by

$$d\mathcal{S}(\vec{x})\,\vec{h} = \left\langle \frac{\delta \mathcal{S}}{\delta \vec{x}}, \vec{h} \right\rangle.$$

Remark 3.8. Considering the scalar case $x : [a, b] \to \mathbb{R}$ for simplicity, and taking $h(t) = \delta_{t_0}(t)$, where $\delta_{t_0}(t) = \delta(t - t_0)$ is the delta function supported at t_0, we have formally that

$$\frac{\delta \mathcal{S}}{\delta x(t_0)} = \frac{d}{d\varepsilon} \mathcal{S}(x + \varepsilon \delta_{t_0}) \Big|_{\varepsilon=0}.$$

Thus, we may interpret the value of the functional derivative $\delta\mathcal{S}/\delta x$ at t as describing the change in the values of the functional $\mathcal{S}(x)$ due to changes in the function x at the point t.

1.4. Examples from mechanics

Let us return to the examples considered in Section 1.

Example 3.9. The action for the one-dimensional oscillator in Example 3.1 is

$$\mathcal{S}(x) = \int_a^b \left\{ \frac{1}{2} m \dot{x}^2 - V(x) \right\} dt,$$

and its variational derivative is

$$\frac{\delta \mathcal{S}}{\delta x} = -\left[m\ddot{x} + V'(x) \right].$$

Example 3.10. The potential energy $V : \mathbb{R}^3 \setminus \{0\} \to \mathbb{R}$ for a central inverse-square force is given by

$$V(\vec{x}) = -\frac{GMm}{|\vec{x}|}.$$

The action of a trajectory $\vec{x} : [a, b] \to \mathbb{R}^3$ is

$$\mathcal{S}(\vec{x}) = \int_a^b \left\{ \frac{1}{2} m \left| \dot{\vec{x}} \right|^2 + \frac{GMm}{|\vec{x}|} \right\} dt.$$

Example 3.11. The action for the n-body problem in Example 3.3 is

$$\mathcal{S}(\vec{x}_1, \vec{x}_2, \ldots, \vec{x_n}) = \int_a^b \left\{ \frac{1}{2} \sum_{i=1}^n m_i \left| \dot{\vec{x}}_i \right|^2 + \frac{1}{2} \sum_{i,j=1}^n \frac{G m_i m_j}{|\vec{x}_i - \vec{x}_j|} \right\} dt.$$

The equations of motion are obtained from the requirement that $\mathcal{S}$ is stationary with respect to independent variations of $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$.

Example 3.12. The configuration of a particle may be described by a point in some other manifold than $\mathbb{R}^n$. For example, consider a pendulum of length ℓ and mass m in a gravitational field with acceleration g. We may describe its configuration by an angle $\theta \in \mathbb{T}$. The action is

$$\mathcal{S} = \int_a^b \left\{ \frac{1}{2} m \ell^2 \dot{\theta}^2 - mg\ell(1 - \cos\theta) \right\} dt,$$

and the corresponding equation of motion is the pendulum equation

$$\ell\ddot{\theta} + g \sin\theta = 0.$$

The following example connects mechanics and the calculus of variations with Riemannian geometry.

Example 3.13. Consider a particle moving freely on a Riemannian manifold M with metric g. If $x = \left(x^1, x^2, \ldots, x^n\right)$ are local coordinates on M, then the arclength ds on M is given by

$$ds^2 = g_{ij}(x)dx^i dx^j$$

where g_{ij} are the metric components. The metric is required to be symmetric, so $g_{ij} = g_{ji}$, and non-singular. We use the summation convention, meaning that repeated upper and lower indices are summed from 1 to n. A trajectory $\gamma : [a, b] \to M$ has kinetic energy

$$T(\gamma, \dot{\gamma}) = \frac{1}{2} g_{ij}(\gamma)\dot{\gamma}^i \dot{\gamma}^j.$$

The corresponding action is

$$\mathcal{S} = \frac{1}{2} \int_a^b g_{ij}(\gamma)\dot{\gamma}^i \dot{\gamma}^j \, dt.$$

The principle of stationary action leads to the equation

$$g_{ij}(\gamma)\ddot{\gamma}^j + \Gamma_{jki}(\gamma)\dot{\gamma}^j \dot{\gamma}^k = 0 \qquad i = 1, 2, \ldots, n$$

where the connection coefficients, or Christoffel symbols, Γ_{jki} are defined by

$$\Gamma_{jki} = \frac{1}{2}\left(\frac{\partial g_{ij}}{\partial x^k} + \frac{\partial g_{ik}}{\partial x^j} - \frac{\partial g_{jk}}{\partial x^i}\right).$$

Since the metric is invertible, we may solve this equation for $\ddot{\gamma}$ to get

$$\ddot{\gamma}^i + \Gamma^i_{jk}(\gamma)\dot{\gamma}^j \dot{\gamma}^k = 0 \qquad i = 1, 2, \ldots, n \tag{3.6}$$

where

$$\Gamma^i_{jk} = g^{ip}\Gamma_{jkp}$$

and g^{ij} denotes the components of the inverse matrix of g_{ij} such that

$$g^{ij}g_{jk} = \delta^i_k.$$

The solutions of the second-order system of ODEs (3.6) are the geodesics of the manifold.

2. The Euler-Lagrange equation

In the mechanical problems considered above, the Lagrangian is a quadratic function of the velocity. Here, we consider Lagrangians with a more general dependence on the derivative.

Let $\mathcal{F}$ be a functional of scalar-valued functions $u : [a, b] \to \mathbb{R}$ of the form

$$\begin{aligned} &\mathcal{F}(u) = \int_a^b F\left(x, u(x), u'(x)\right) \, dx, \\ &F : [a, b] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \end{aligned} \tag{3.7}$$

where F is a smooth function.

It is convenient to use the same notation for the variables

$$(x, u, u') \in [a, b] \times \mathbb{R} \times \mathbb{R}$$

on which F depends and the functions $u(x)$, $u'(x)$. We denote the partial derivatives of $F(x, u, u')$ by

$$F_x = \left.\frac{\partial F}{\partial x}\right|_{u,u'}, \qquad F_u = \left.\frac{\partial F}{\partial u}\right|_{x,u'}, \qquad F_{u'} = \left.\frac{\partial F}{\partial u'}\right|_{x,u}.$$

If $h : [a, b] \to \mathbb{R}$ is a smooth function that vanishes at $x = a, b$, then

$$\begin{aligned} d\mathcal{F}(\vec{u})\, h &= \frac{d}{d\varepsilon} \int_a^b F\left(x, u(x) + \varepsilon h(x), u'(x) + \varepsilon h'(x)\right)\, dx \Bigg|_{\varepsilon=0} \\ &= \int_a^b \left\{ F_u\left(x, u(x), u'(x)\right) h(x) + F_{u'}\left(x, u(x), u'(x)\right) h'(x) \right\}\, dx. \end{aligned} \tag{3.8}$$

It follows that a necessary condition for a C^1-function $u(x)$ to be a stationary point of (3.7) in a space of functions with given values at the endpoints is that

$$\int_a^b \left\{ F_u\left(x, u(x), u'(x)\right) h(x) + F_{u'}\left(x, u(x), u'(x)\right) h'(x) \right\}\, dx = 0 \tag{3.9}$$

for all smooth functions $h(x)$ that vanish at $x = a, b$.

If the function u in (3.8) is C^2, then we may integrate by parts to get

$$d\mathcal{F}(\vec{u})\, h = \int_a^b \left\{ F_u\left(x, u(x), u'(x)\right) - \frac{d}{dx} F_{u'}\left(x, u(x), u'(x)\right) \right\} h(x)\, dx.$$

It follows that the variational derivative of $\mathcal{F}$ is given by

$$\frac{\delta \mathcal{F}}{\delta u} = -\frac{d}{dx} F_{u'}\left(x, u, u'\right) + F_u\left(x, u, u'\right).$$

Moreover, if a C^2-function $u(x)$ is a stationary point of $\mathcal{F}$, then it must satisfies the ODE

$$-\frac{d}{dx} F_{u'}\left(x, u, u'\right) + F_u\left(x, u, u'\right) = 0. \tag{3.10}$$

Equation (3.10) is the *Euler-Lagrange equation* associated with the functional (3.7). It is a necessary, but not sufficient, condition that any smooth minimizer of (3.7) must satisfy. Equation (3.9) is the weak form of (3.10); it is satisfied by any C^1-minimizer (or, more generally, by any minimizer that belongs to a suitable Sobolev space $W^{1,p}(a, b)$).

Note that d/dx in (3.10) is the total derivative with respect to x, meaning that the derivative is taken after the substitution of the functions $u(x)$ and $u'(x)$ into the arguments of F. Thus,

$$\frac{d}{dx} f\left(x, u, u'\right) = f_x\left(x, u, u'\right) + f_u\left(x, u, u'\right) u' + f_{u'}\left(x, u, u'\right) u''.$$

The coefficient of u'' in (3.10) is equal to $F_{u'u'}$. The ODE is therefore of second order provided that

$$F_{u'u'}\left(x, u, u'\right) \neq 0.$$

The derivation of the Euler-Lagrange equation extends straightforwardly to Lagrangians that depend on higher derivatives and to systems. For example, the Euler-Lagrange equation for the scalar functional

$$\mathcal{F}(u) = \int_a^b F\left(x, u(x), u'(x), u''(x)\right)\, dx,$$

where $F : [a, b] \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, is

$$\frac{d^2}{dx^2} F_{u''} - \frac{d}{dx} F_{u'} + F_u = 0.$$

This is a forth-order ODE if $F_{u''u''} \neq 0$.

The Euler-Lagrange equation for a vector functional

$$\mathcal{F}(\vec{u}) = \int_a^b F(x, \vec{u}(x), \vec{u}'(x))\, dx,$$

where $F : [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, is

$$-\frac{d}{dx} F_{u_i'} + F_{u_i} = 0 \qquad \text{for } i = 1, 2, \ldots, n.$$

This is an $n \times n$ system of ODEs for $\vec{u} = (u_1, u_2, \ldots, u_n)$. The system is second-order if the $n \times n$ matrix with components $F_{u_i' u_j'}$ is invertible.

The extension to functionals that involve more than one independent variable is less straightforward, and some examples will be considered below. In that case, the Euler-Lagrange equation is a PDE.

The question of whether a solution of the Euler-Lagrange equation is an extreme point of the functional is quite subtle even in the one-dimensional case. For example, the application of a second-derivative test, familiar from calculus for functions on finite-dimensional spaces, is not entirely straightforward. We will not discuss these questions here; see [**11**], for example, for more information.

3. Newton's problem of minimal resistance

> If in a rare medium, consisting of equal particles freely disposed at equal distance from each other, a globe and a cylinder described on equal diameter move with equal velocities in the direction of the axis of the cylinder, the resistance of the globe will be half as great as that of the cylinder. ... I reckon that this proposition will not be without application in the building of ships.[2]

Many variational problems arise from optimization problems in which we seek to minimize (or maximize) some functional. We consider here a problem proposed and solved by Newton (1685) of finding the shape of a body with minimal resistance in a rarified gas. This was one of the first problems in the calculus of variations to be solved.

3.1. Derivation of Newton's resistance functional

Following Newton, let us imagine that the gas is composed of uniformly distributed, non-interacting particles that reflect elastically off the body. We suppose that the particles have number-density n, mass m, and constant velocity v the downward z-direction, in a frame of reference moving with the body.

We assume that the body is cylindrically symmetric with a maximum radius of a and height h. We write the equation of the body surface in cylindrical polar coordinates as $z = u(r)$, where $0 \le r \le a$ and

$$u(0) = h, \qquad u(a) = 0.$$

Let $\theta(r)$ denote the angle of the tangent line to the r-axis of this curve at the point $(r, u(r))$. Since the angle of reflection of a particle off the body is equal to the angle of incidence, $\pi/2 - \theta$, the reflected particle path makes an angle 2θ to the z-axis.

---

[2] I. Newton in *Principia Mathematica*, quoted from [**11**].

The change in momentum of the particle in the z-direction when it reflects off the body is therefore

$$mv\,(1+\cos 2\theta)\,.$$

For example, this is equal to $2mv$ for normal incidence ($\theta = 0$), and 0 for grazing incidence ($\theta = \pi/2$).

The number of particles per unit time, per unit distance in the radial direction that hit the body is equal to

$$2\pi nvr.$$

Note that $[2\pi nvr] = (1/L^3)\cdot(L/T)\cdot(L) = 1/(LT)$ as it should.

The rate at which the particles transfer momentum to the body per unit time, which is equal to force F exerted by the gas on the body, is given by

$$F = 2\pi nmv^2 \int_0^a r\,(1+\cos 2\theta)\;dr.$$

Using the fact that $\tan\theta = u'$ to eliminate θ, we get that the resistance force on a profile $z = u(r)$ is given by

$$F = 4\pi nma^2v^2\mathcal{F}(u),$$

where the resistance functional $\mathcal{F}$ is defined by

$$\mathcal{F}(u) = \frac{1}{a^2}\int_0^a \frac{r}{1+[u'(r)]^2}\,dr. \tag{3.11}$$

Introducing dimensionless variables $\tilde{r} = r/a$, $\tilde{u} = u/a$ in (3.11), and dropping the tildes, we get the nondimensionalized resistance functional

$$\mathcal{F}(u) = \int_0^1 \frac{r}{1+[u'(r)]^2}\,dr. \tag{3.12}$$

As we will see, this resistance functional does not provide the the most convincing physical results, although it has been used as a model for rarified flows and hypersonic flows. It is nevertheless remarkable that Newton was able to formulate and solve this problem long before a systematic development of the theory of fluid mechanics.

3.2. Resistances of some simple shapes

To see how the resistance functional $\mathcal{F}$ in (3.11) behaves and formulate an appropriate optimization problem for it, let us consider some examples. Clearly, we have $0 < \mathcal{F}(u) \le 1/2$ for any $u : [a,b] \to \mathbb{R}$.

Example 3.14. For a vertical cylinder of radius a, we have $u(r) = h$ for $0 \le r < a$ and $u(a) = 0$. The integrand in the functional (3.11) is small when u' is large, so we can approximate this discontinuous function by smooth functions whose resistance is arbitrarily close to the resistance of the cylinder. Setting $u' = 0$ in (3.11), we get $\mathcal{F} = 1/2$. Thus, a blunt cylinder has the maximum possible resistance. The resistance is independent of the cylinder height h, since the gas particles graze the sides of the cylinder and exert no force upon it.

Example 3.15. For a sphere, with $r^2 + z^2 = a^2$ and $u(r) = \sqrt{a^2 - r^2}$, we get $\mathcal{F} = 1/4$. As Newton observed, this is half the resistance of the cylinder. More generally, consider an ellipsoid of radius a and height h, with aspect ratio

$$M = \frac{h}{a}, \tag{3.13}$$

and equation

$$\frac{r^2}{a^2} + \frac{z^2}{h^2} = 1, \qquad u(r) = M\sqrt{a^2 - r^2}.$$

Using this expression for u in (3.11), and assuming that $M \neq 1$, we get the resistance

$$\mathcal{F}(u) = \frac{M^2 \log M^2 - \left(M^2 - 1\right)}{2\left(M^2 - 1\right)^2}.$$

The limit of this expression as $M \to 0$ is equal to $1/2$, the resistance of the cylinder, and the limit as $M \to 1$ is $1/4$, the resistance of the sphere. As $M \to \infty$, the resistance approaches zero. Thus, the resistance becomes arbitrarily small for a sufficiently tall, thin ellipsoid, and there is no profile that minimizes the resistance without a constraint on the aspect ratio.

Example 3.16. The equation of a circular cone with base a and height h is $z = u(r)$ with $u(r) = M(a - r)$, where M is given by (3.13) as before. In this case $u' = M$ is constant, and

$$\mathcal{F}(u) = \frac{1}{2\left(1 + M^2\right)}$$

As $M \to 0$, the resistance approaches $1/2$, and as $M \to \infty$, the resistance approaches 0.

Example 3.17. Suppose that $u_n(r)$ consists of $(n + 1/2)$ 'tent' functions of height h and base $2b_n$ where

$$b_n = \frac{a}{2n+1}.$$

Then, except at the 'corners,' we have $|u_n'| = h/b_n$, and therefore

$$\mathcal{F}(u_n) = \frac{1}{2\left[1 + (2n+1)^2 M^2\right]}.$$

As before, we can approximate this piecewise smooth function by smooth functions with an arbitrarily small increase in the resistance. Thus, $\mathcal{F}(u_n) \to 0$ as $n \to \infty$, even though $0 \leq u_n(r) \leq h$ and the heights of the bodies are uniformly bounded. To eliminate this kind of oscillatory behavior, which would lead to multiple impacts of particles on the body contrary to what is assumed in the derivation of the resistance formula, we will impose the reasonable requirement that $u'(r) \leq 0$ for $0 \leq r \leq a$.

3.3. The variational problem

We fix the aspect ratio $M > 0$, and seek to minimize $\mathcal{F}$ over the space of functions

$$X_M = \left\{u \in W^{1,\infty}(0,1) : [0,1] \to \mathbb{R} \mid u(0) = M, \quad u(1) = 0, \quad u'(r) \leq 0\right\}.$$

Here, $W^{1,\infty}(0,1)$ denotes the Sobolev space of functions whose weak, or distributional, derivative is a bounded function $u' \in L^\infty(0,1)$. Equivalently, this means that u is Lipschitz continuous with $|u(x) - u(y)| \leq M|x - y|$, where $M = \|u\|_\infty$. We could minimize $\mathcal{F}$ over the larger space $W^{1,1}(0,1)$ of absolutely continuous functions with $u' \in L^1(0,1)$, and get the same result. As we shall see, however, the smaller space $C^1[0,1]$ of continuously differentiable functions would not be adequate because the minimizer has a 'corner' and is not continuously differentiable.

Also note that, as the examples above illustrate, it is necessary to impose a constraint, such as $u' \leq 0$, on the admissible functions, otherwise (as pointed out by Legendre in 1788) we could make the resistance as small as we wish by taking

profiles with rapid 'zig-zags' and large slopes, although the infimum $\mathcal{F} = 0$ is not attained for any profile.

The functional (3.12) is a pathological one from the perspective of the general theory of the calculus of variations. First, it is not coercive, because

$$\frac{r}{1 + [u']^2} \to 0 \qquad \text{as } |u'| \to \infty.$$

As a result, minimizing sequences need not be bounded, and, in the absence of constraints, minimizers can 'escape' to infinity. Second, it is not convex. A function $\mathcal{F} : X \to \mathbb{R}$ on a real vector space X is convex if, for all $u, v \in X$ and $\lambda \in [0, 1]$,

$$\mathcal{F}\left(\lambda u + (1 - \lambda)v\right) \leq \lambda \mathcal{F}(u) + (1 - \lambda)\mathcal{F}(v).$$

In general, convex functions have good lower semicontinuity properties and convex minimization problems are typically well-behaved. The behavior of non-convex optimization problems can be much nastier.

3.4. The Euler-Lagrange equation

The Euler-Lagrange equation for (3.12) is

$$\frac{d}{dr}\left\{ \frac{ru'}{\left[1 + (u')^2\right]^2} \right\} = 0.$$

Since the Lagrangian is independent of u, this has an immediate first integral,

$$ru' = -c\left[1 + (u')^2\right]^2 \tag{3.14}$$

where $c \geq 0$ is a constant of integration.

If $c = 0$ in (3.14), then we get $u' = 0$, or $u =$ constant. This solution corresponds to the cylinder with maximum resistance $1/2$. The maximum is not attained, however, within the class absolutely continuous functions $u \in X_M$, since for such functions if u' is zero almost everywhere with respect to Lebesgue measure, then u is constant, and it cannot satisfy both boundary conditions $u(0) = M$, $u(1) = 0$.

If $c > 0$ in (3.14), then it is convenient to parametrize the solution curve by $p = u' < 0$. From (3.14), the radial coordinate r is given in terms of p by

$$r = -\frac{c\left(1 + p^2\right)^2}{p}. \tag{3.15}$$

Using this equation to express dr in terms of dp in the integral

$$u = \int p\, dr,$$

and evaluating the result, we get

$$u = u_0 - c\left(-\log|p| + p^2 + \frac{3}{4}p^4\right), \tag{3.16}$$

where u_0 is a constant of integration.

From (3.15), we see that the minimum value of $r(p)$ for $p < 0$ is

$$r_0 = \frac{16\sqrt{3}c}{9}$$

at $p = -1/\sqrt{3}$. Thus, although this solution minimizes the resistance, we cannot use it over the whole interval $0 \le r \le 1$, only for $r_0 \le r \le 1$. In the remaining part of the interval, we use $u =$ constant, and we obtain the lowest global resistance by placing the blunt part of the body around the nose $r = 0$, where it contributes least to the area and resistance.

While this plausibility argument seems reasonable, it is not entirely convincing, since the flat nose locally maximizes the resistance, and it is far from a proof. Nevertheless, with additional work, it is possible to prove that it does give the correct solution $u \in X_M$ with minimal resistance.

This minimizing solution has the form

$$u(r) = \begin{cases} M & \text{for } 0 \le r \le r_0, \\ u_0 - c\left(-\log|p| + p^2 + \frac{3}{4}p^4\right) & \text{for } p_1 \le p \le -1/\sqrt{3}, \end{cases}$$

where $r(p_1) = 1$.

Imposing continuity of the solution at $r = r_0$, $p = 1/\sqrt{3}$ and the boundary condition $u(1) = 0$, with $p = p_1$, we get

$$\begin{aligned} M &= u_0 - c\left(\log\sqrt{3} + \frac{5}{12}\right), \\ p_1 &= -c\left(1 + p_1^2\right)^2, \\ u_0 &= c\left(-\log|p_1| + p_1^2 + \frac{3}{4}p_1^4\right). \end{aligned}$$

Eliminating u_0, we may write the solution as

$$u(r) = M - c\left(p^2 + \frac{3}{4}p^4 - \log\left|\sqrt{3}p\right| - \frac{5}{12}\right)$$

for $p_1 \le p \le -1/\sqrt{3}$, where

$$M = c\left(p_1^2 + \frac{3}{4}p_1^4 - \log\left|\sqrt{3}p_1\right| - \frac{5}{12}\right), \qquad p_1 = -c\left(1 + p_1^2\right)^2.$$

Thus, p_1 is the solution of

$$\frac{p_1\left(\log\left|\sqrt{3}p_1\right| - p_1^2 - \frac{3}{4}p_1^4 + \frac{5}{12}\right)}{\left(1 + p_1^2\right)^2} = M,$$

and r_0 is given in terms of p_1 by

$$r_0 = -\frac{16\sqrt{3}p_1}{9\left(1 + p_1^2\right)^2}.$$

Denoting by

$$C_0 = 2\int_0^1 \frac{r}{1 + (u')^2}\, dr$$

the ratio of the minimal resistance to the maximal resistance of a cylinder, one gets the numerical values shown below [**12**]. Moreover, one can show that

$$r_0 \sim \frac{27}{16}\frac{1}{M^3}, \quad C_0 \sim \frac{27}{32}\frac{1}{M^2} \qquad \text{as } M \to \infty.$$

Thus, as the aspect ratio increases, the radius of the blunt nose decreases and the total resistance of the body approaches zero.

	$M=1$	$M=2$	$M=3$	$M=4$
r_0	0.35	0.12	0.048	0.023
C_0	0.37	0.16	0.0082	0.0049

3.5. Non-radially symmetric solutions

The radially symmetric problem may be generalized to a two-dimensional, non-radially symmetric problem as follows. Suppose that $\Omega \subset \mathbb{R}^2$ is a given domain (a bounded, open, connected set). Find a bounded, nonegative convex function $u : \Omega \to \mathbb{R}$ that minimizes

$$\mathcal{F}(u) = \int_\Omega \frac{1}{1 + |\nabla u|^2}\, dxdy.$$

In this case, the shape of the body is given by $z = u(x, y)$.

In the discussion above, we obtained the minimizer among radially symmetric bodies when Ω is a disc D. It might seem natural to suppose that this radially symmetric solution minimizes the resistance among non-radially symmetric admissible functions $u : D \to \mathbb{R}$. It is interesting to note that this is not true. Brock, Ferroni, and Kawohl (1996) showed that there are non-radially symmetric convex functions on the disc that give a lower resistance than the radial solution found above.

4. Constrained variational principles

It often occurs that we want to minimize a functional subject to a constraint. Constraints can take many forms. First, consider the minimization of a functional

$$\mathcal{F}(u) = \int_a^b F\left(x, u, u'\right)\, dx,$$

over functions such that $u(a) = 0$, $u(b) = 0$, subject to an integral constraint of the form

$$\mathcal{G} = \int_a^b G\left(x, u, u'\right)\, dx.$$

Variational problems with integral constraints are called isoperimetric problems after the prototypical problem of finding the curve (a circle) that encloses the maximum area subject to the constraint that its length is fixed.[3]

We may solve this problem by introducing a Lagrange multiplier $\lambda \in \mathbb{R}$ and seeking stationary points of the unconstrained functional

$$\mathcal{F}(u) - \lambda\mathcal{G}(u) = \int_a^b \left\{F\left(x, u, u'\right) - \lambda G\left(x, u, u'\right)\right\}\, dx.$$

The condition that this functional is stationary with respect to λ implies that $\mathcal{G}(u) = 0$, so a stationary point satisfies the constraint.

The Euler-Lagrange equation for stationarity of the functional with respect to variations in u is

$$-\frac{d}{dx}F_{u'}\left(x, u, u'\right) + F_u\left(x, u, u'\right) = \lambda\left[-\frac{d}{dx}G_{u'}\left(x, u, u'\right) + G_u\left(x, u, u'\right)\right].$$

[3]According to Virgil's Aeneid, Dido was given as much land as she could enclose with an ox hide to found the city of Carthage. She cut the hide into a thin strip, and used it to enclose a large circular hill.

In principle, we solve this problem for $u(x)$ and λ subject to the boundary conditions $u(a) = 0$, $u(b) = 0$ and the constraint $\mathcal{G}(u) = 0$.

4.1. Eigenvalue problems

Consider the following Rayleigh quotient

$$\mathcal{Q}(u) = \frac{\int_a^b \left\{ p(x)u'^2 + q(x)u^2 \right\} \, dx}{\int_a^b u^2 \, dx}$$

where $p(x)$, $q(x)$ are given coefficient functions.

Since $\mathcal{Q}(u)$ is homogeneous in u, the minimization of $\mathcal{Q}(u)$ over nonzero functions u is equivalent to the minimization of the numerator subject to the constraint that the denominator is equal to one; or, in other words, to the minimization of $\mathcal{F}(u)$ subject to the constraint $\mathcal{G}(u) = 0$ where

$$\mathcal{F}(u) = \frac{1}{2} \int_a^b \left\{ p(x)u'^2 + q(x)u^2 \right\} \, dx, \qquad \mathcal{G}(u) = \frac{1}{2} \left\{ \int_a^b u^2 \, dx - 1 \right\}.$$

The corresponding Euler-Lagrange equation for the stationarity of $\mathcal{F}(u) - \lambda \mathcal{G}(u)$ with respect to u is

$$-\left[p(x)u' \right]' + q(x)u = \lambda u.$$

This is a Sturm-Liouville eigenvalue problem in which the Lagrange multiplier λ is an eigenvalue.

5. Elastic rods

As an example of the use of constrained variational principles, we will derive equations for the equilibria of an inextensible elastic rod and describe some applications.

Consider a thin, inextensible elastic rod that resists bending. Suppose that the cross-sections of the rod are isotropic and that we can ignore any twisting. We model the spatial configuration of the rod by a curve $\vec{r}(s)$,

$$\vec{r} : [a, b] \to \mathbb{R}^3,$$

where it is convenient to parametrize the curve by arclength $a \leq s \leq b$.

We can model the twisting of a rod by introducing additional vectors that describe the orientation of its cross-section, leading to the Kirchoff and Cosserat theories [**4**], but we will not consider such generalizations here.

5.1. Kinematics

We introduce an orthonormal frame of vectors $\{\vec{t}, \vec{n}, \vec{b}\}$ along the curve, consisting of the unit tangent, normal and binormal vectors, respectively. We have $\vec{t} = \vec{r}'$ and $\vec{b} = \vec{t} \times \vec{n}$. According to the the Frenet-Serret formulas, these vectors satisfy

$$\vec{t}' = \kappa \vec{n}, \qquad \vec{n}' = -\kappa \vec{t} + \tau \vec{b}, \qquad \vec{b}' = -\tau \vec{n}$$

where $\kappa(s)$ is the curvature and $\tau(s)$ is the torsion of the curve.

These equations may also be written as

$$\begin{pmatrix} \vec{t} \\ \vec{n} \\ \vec{b} \end{pmatrix}' = \begin{pmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{pmatrix} \begin{pmatrix} \vec{t} \\ \vec{n} \\ \vec{b} \end{pmatrix}.$$

The skew-symmetric matrix on the right-hand side is the infinitesimal generator of the rotations of the orthonormal frame $\{\vec{t}, \vec{n}, \vec{b}\}$ as it is transported along the curve.

5.2. A variational principle

We will derive equilibrium equations for the configuration of the rod from the condition that they minimize the energy.

We assume that the energy density of a rod configuration is proportional to the square of its curvature. This constitutive equation, and the model of a rod as an 'elastic line,' or *elastica*, was introduced and developed by James Bernoulli[4] (1694), Daniel Bernoulli (1728), and Euler (1727, 1732).

The curvature is given by $\kappa^2 = \vec{t}' \cdot \vec{t}'$, so the total energy of the rod is given by

$$\mathcal{E}(\vec{r}) = \int_a^b \frac{1}{2} J\, \vec{r}'' \cdot \vec{r}''\, ds, \tag{3.17}$$

where the material function $J : [a, b] \to \mathbb{R}$ gives the proportionality between the square of the curvature and the energy density due to bending.

Equations for the equilibrium configuration of the rod follow by minimizing the energy (3.17) subject to the constraint that s is arclength, meaning that

$$\vec{r}' \cdot \vec{r}' = 1.$$

This constraint is a pointwise constraint, rather than an integral, so we impose it by introducing a function $\lambda : [a, b] \to \mathbb{R}$ as a Lagrange multiplier, and seeking stationary points of the functional

$$\mathcal{F}(\vec{r}, \lambda) = \int_a^b \frac{1}{2} \left\{ J\, \vec{r}'' \cdot \vec{r}'' - \lambda \left(\vec{r}' \cdot \vec{r}' - 1 \right) \right\} ds.$$

The Euler-Lagrange equation obtained by varying $\vec{r}$ is

$$\left(J \vec{r}'' \right)'' + \left(\lambda \vec{r}' \right)' = 0, \tag{3.18}$$

while we recover the constraint by varying λ. Integrating (3.18) once, and writing $\vec{r}' = \vec{t}$, we get

$$\left(J \vec{t}' \right)' + \lambda \vec{t} = \vec{F} \tag{3.19}$$

where $\vec{F}$ is a constant vector of integration. It corresponds to the contact force exerted by one part of the rod on another, which is constant in an inextensible rod which is not acted on by an external force.

We obtain an expression for the Lagrange multiplier λ by imposing the constraint that $\vec{t}$ is a unit vector on solutions of (3.19). Taking the inner product of (3.19) with $\vec{t}$, and rewriting the result, we get

$$\left(J \vec{t} \cdot \vec{t}' \right)' - J \vec{t}' \cdot \vec{t}' + \lambda \vec{t} \cdot \vec{t} = \vec{F} \cdot \vec{t}.$$

Using $\vec{t} \cdot \vec{t} = 1$ and $\vec{t} \cdot \vec{t}' = 0$ in this equation, we get

$$\lambda = \vec{F} \cdot \vec{t} + J \vec{t}' \cdot \vec{t}'.$$

---

[4]There were a lot of Bernoulli's. The main ones were the older brother James (1654-1705), the younger brother Johann (1667-1748), and Johann's son Daniel (1700-1782). James and Johann has a prolonged feud over the priority of their mathematical results, and, after James died, Johann became jealous of his son Daniel's work, in particular on Bernoulli's law in hydrodynamics.

Thus, (3.19) may be written as

$$\left(J\vec{t}'\right)' + J\kappa^2\vec{t} = \vec{F} - \left(\vec{F}\cdot\vec{t}\right)\vec{t}, \qquad \kappa^2 = \vec{t}'\cdot\vec{t}'. \tag{3.20}$$

Equation (3.20) is a second order ODE for the tangent vector $\vec{t}(s)$. We supplement it with suitable boundary conditions at the ends of rod. For example, if the ends are fully clamped, we specify the directions $\vec{t}(a)$, $\vec{t}(b)$ of the rod at each endpoint. Given a solution for $\vec{t}$, we may then recover the position of the rod by integrating the equation $\vec{r}' = \vec{t}$. Note that, in this case, we cannot expect to also specify the position of both endpoints. In general, the issue of what boundary conditions to use in rod theories is somewhat subtle (see [**4**] for further discussion).

Taking the cross product of (3.20) with $\vec{t}$, and using the fact that $\vec{t}\times\vec{t}' = \kappa\vec{b}$, we get

$$\vec{m}' = \vec{t}\times\vec{F}, \qquad \text{where } \vec{m} = J\kappa\vec{b}.$$

This equation expresses a balance of moments in the rod due to the constant contact force $\vec{F}$ and a contact couple $\vec{m}$. The couple is proportional to the curvature, as proposed by Bernoulli and Euler, corresponding to the constitutive assumption that the energy density is a quadratic function of the curvature Thus, we obtain the same equations from the Euler-Lagrange equations of the variational principle as we would by balancing the forces and moments acting on the rod.

5.3. Dimensional considerations

From (3.17), the material function J has the dimension of energy$\cdot$length. It is often written as $J = EI$ where E is Young's modulus for the elastic material making up the rod, and I is the moment of inertia of a cross-section.

Young's modulus gives the ratio of tensile stress to tensile strain in an elastic solid. Strain, which measures a deformed length to an undeformed length, is dimensionless, so E has the dimension of stress, force/area, meaning that

$$[E] = \frac{M}{LT^2}.$$

For example, the Young's modulus of steel is approximately $200\,\mathrm{kN/mm^2}$.

The moment of inertia, in this context, is a second area moment of the rod cross-section, and has the dimension of L^4. The term 'moment of inertia' is also used to describe the relationship between angular velocity and angular momentum for a rotating rigid body; the moment of inertia here corresponds to this notion with mass replaced by area.

Explicitly, we define the components of a second-order, area-moment tensor of a region $\Omega\subset\mathbb{R}^2$ in the plane, with Cartesian coordinates x_i, $i = 1, 2$, by

$$I_{ij} = \int_\Omega x_i x_j\, dA.$$

In general, this symmetric, positive-definite tensor has two positive real eigenvalues, corresponding to the moments of inertia about the principal axes defined by the corresponding eigenvectors. If these eienvalues coincide, then we get the the isotropic case with $I_{ij} = I\delta_{ij}$ where I is the moment of inertia. For example, if Ω is a disc of radius a, then $I = \pi a^4/4$.

Thus,

$$[EI] = \frac{ML^2}{T^2}\cdot L,$$

consistent with the dimension of J. In general, J may depend upon s, for example because the cross-sectional area of the rod, and therefore moment of inertia, varies along its length.

5.4. The persistence length of DNA

An interesting application of rod theories is to the modeling of polymers whose molecular chains resist bending, such as DNA. A statistical mechanics of flexible polymers may be derived by supposing that the polymer chain undergoes a random walk due to thermal fluctuations. Such polymers typically coil up because there are more coiled configurations that straight ones, so coiling is entropically favored.

If a polymer has elastic rigidity, then the increase in entropy that favors its coiling is opposed by the bending energy required to coil. As a result, the tangent vector of the polymer chain is highly correlated over distances short enough that significant bending energies are required to change its direction, while it is decorrelated over much longer distances. A typical lengthscale over which the tangent vector is correlated is called the *persistence length* of the polymer.

According to statistical mechanics, the probability that a system at absolute temperature T has a specific configuration with energy E is proportional to

$$e^{-E/kT} \tag{3.21}$$

where k is Boltzmann's constant. Boltzmann's constant has the approximate value $k = 1.38 \times 10^{-23}\,\mathrm{J}K^{-1}$. The quantity kT is an order of magnitude for the random thermal energy of a single microscopic degree of freedom at temperature T.

The bending energy of an elastic rod is set by the coefficent J in (3.17), with dimension energy $\cdot$ length. Thus, the quantity

$$A = \frac{J}{kT}$$

is a lengthscale over which thermal and bending energies are comparable, and it provides a measure of the persistence length. For DNA, a typical value of this length at standard conditions is $A \approx 50\,\mathrm{nm}$, or about 150 base pairs of the double helix.

The statistical mechanics of an elastica, or 'worm-like chain,' may be described, formally at least, in terms of path integrals (integrals over an infinite-dimensional space of functions). The expected value $\mathbf{E}\left[\mathcal{F}\left(\vec{r}\right)\right]$ of some functional $\mathcal{F}\left(\vec{r}\right)$ of the elastica configuration is given by

$$\mathbf{E}\left[\mathcal{F}\left(\vec{r}\right)\right] = \frac{1}{Z}\int \mathcal{F}\left(\vec{r}\right) e^{-\mathcal{E}(\vec{r})/kT}\, D\vec{r},$$

where the right-hand side is a path integral over a path space of configurations $\vec{r}(s)$ using the Boltzmann factor (3.21) and the elastica energy (3.17). The factor Z is inserted to normalize the Boltzmann distribution to a probability distribution.

These path integrals are difficult to evaluate in general, but in some cases the energy functional may be approximated by a quadratic functional, and the resulting (infinite-dimensional) Gaussian integrals may be evaluated exactly. This leads to results which are in reasonable agreement with the experimentally observed properties of DNA [**36**]. One can also include other effects in the model, such as the twisting energy of DNA.

6. Buckling and bifurcation theory

Let us consider planar deformations of an elastic rod of length L. In this case, we may write

$$\vec{t} = (\cos\theta, \sin\theta)$$

in (3.20), where $\theta(s)$ is the angle of the rod to the x-axis. We assume that the rod is uniform, so that $J = EI$ is constant, and that the force $\vec{F} = (F, 0)$ in the rod is directed along the x-axis, with $F > 0$, corresponding to a compression.

With these assumptions, equation (3.20) reduces to a scalar ODE

$$EI\theta'' + F\sin\theta = 0.$$

This ODE is the Euler-Lagrange equation of the functional

$$\mathcal{E}(\theta) = \int_0^L \left\{ \frac{1}{2} EI\,(\theta')^2 - F\,(1 - \cos\theta) \right\} ds$$

The first term is the bending energy of the rod, and the second term is the work done by the force in shortening the length of the rod in the x-direction.

This equation is identical in form to the pendulum equation. Here, however, the independent variable is arclength, rather than time, and we will impose boundary conditions, not initial conditions, on the solutions.

Let us suppose that the ends of the rod at $s = 0$, $s = L$ are horizontally clamped, so that $\theta(0) = 0$, $\theta(L) = 0$. Introducing a dimensionless arclength variable $\tilde{s} = s/L$, and dropping the tildes, we may write this BVP as

$$\theta'' + \lambda\sin\theta = 0, \tag{3.22}$$

$$\theta(0) = 0, \qquad \theta(1) = 0, \tag{3.23}$$

where the dimensionless force parameter $\lambda > 0$ is defined by

$$\lambda = \frac{FL^2}{EI}.$$

This problem was studied by Euler, and is one of the original problems in the bifurcation theory of equilibria.

The problem (3.22)–(3.23) has the trivial solution $\theta = 0$ for any value of λ, corresponding to the unbuckled state of the rod. This is the unique solution when λ is sufficiently small, but other non-trivial solutions bifurcate off the trivial solution as λ increases. This phenomenon corresponds to the buckling of the rod under an increased load.

The problem can be solved explicitly in terms of elliptic functions, as we will show below. First, however, we will obtain solutions by perturbing off the trivial solution. This method is applicable to more complicated problems which cannot be solved exactly.

6.1. The bifurcation equation

To study the bifurcation of non-zero solutions off the zero solution, we first linearize (3.22)–(3.23) about $\theta = 0$. This gives

$$\begin{aligned} &\theta'' + \lambda_0\theta = 0, \\ &\theta(0) = 0, \qquad \theta(1) = 0. \end{aligned} \tag{3.24}$$

We denote the eigenvalue parameter in the linearized problem by λ_0.

Equation (3.24) has a unique solution $\theta = 0$ except when $\lambda_0 = \lambda_0^{(n)}$, where the eigenvalues $\lambda_0^{(n)}$ are given by

$$\lambda_0^{(n)} = n^2\pi^2 \qquad \text{for } n \in \mathbb{N}.$$

The corresponding solutions are then $\theta(s) = A\theta^{(n)}(s)$, where

$$\theta^{(n)}(s) = \sin(n\pi s).$$

The implicit function theorem implies that if $\bar{\lambda}$ is not an eigenvalue of the linearized problem, then the zero solution is the unique solution of the nonlinear problem for (θ, λ) in a small enough neighborhood of $(0, \bar{\lambda})$.

On the other hand, non-trivial solutions can bifurcate off the zero solution at eigenvalues of the linearized problem. We will compute these solutions by expanding the nonlinear problem about an eigenvalue. As we discuss below, this formal computation can be made rigorous by use of a Lyapunov-Schmidt reduction.

Fix $n \in \mathbb{N}$, and let

$$\lambda_0 = n^2\pi^2$$

be the n^{th} eigenvalue. We drop the superscript n to simplify the notation.

We introduce a small parameter ε, and consider values of the eigenvalue parameter λ close to λ_0. We suppose that $\lambda(\varepsilon)$ has the expansion

$$\lambda(\varepsilon) = \lambda_0 + \varepsilon^2\lambda_2 + \dots \qquad \text{as } \varepsilon \to 0, \tag{3.25}$$

where we write ε^2 instead of ε to simplify the subsequent equations.

We look for small-amplitude solutions $\theta(s;\varepsilon)$ of (3.22)–(3.23) with an expansion of the form

$$\theta(s;\varepsilon) = \varepsilon\theta_1(s) + \varepsilon^3\theta_3(s) + \dots \text{as } \varepsilon \to 0. \tag{3.26}$$

Using (3.25) and (3.26) in (3.22)–(3.23), Taylor expanding the result with respect to ε, and equating coefficients of ε and ε^3 to zero, we find that

$$\begin{aligned} &\theta_1'' + \lambda_0\theta_1 = 0, \\ &\theta_1(0) = 0, \qquad \theta_1(1) = 0, \end{aligned} \tag{3.27}$$

$$\begin{aligned} &\theta_3'' + \lambda_0\theta_3 + \lambda_2\theta_1 - \frac{1}{6}\lambda_0\theta_1^3 = 0, \\ &\theta_3(0) = 0, \qquad \theta_3(1) = 0, \end{aligned} \tag{3.28}$$

The solution of (3.27) is

$$\theta_1(s) = A\sin(n\pi s), \tag{3.29}$$

where A is an arbitrary constant of integration.

Equation (3.28) then becomes

$$\begin{aligned} &\theta_3'' + \lambda_0\theta_3 + \lambda_2 A\sin(n\pi s) - \frac{1}{6}\lambda_0 A^3\sin^3(n\pi s) = 0, \\ &\theta_3(0) = 0, \qquad \theta_3(1) = 0, \end{aligned}$$

In general, this equation is not solvable for θ_3. To derive the solvability condition, we multiply the ODE by the eigenfunction $\sin(n\pi s)$ and integrate the result over $0 \le s \le 1$.

Integration by parts, or Green's formula, gives

$$\int_0^1 \sin(n\pi s)\left\{\theta_3{}'' + \lambda_0\theta_3\right\}\,ds - \int_0^1 \left\{\sin(n\pi s)'' + \lambda_0 \sin(n\pi s)\right\}\theta_3\,ds$$
$$= \left[\sin(n\pi s)\,\theta_3{}' - \sin(n\pi s)'\theta_3\right]_0^1.$$

It follows that

$$\int_0^1 \sin(n\pi s)\left\{\theta_3{}'' + \lambda_0\theta_3\right\}\,ds = 0,$$

and hence that

$$\lambda_2 A\int_0^1 \sin^2(n\pi s)\,ds = \frac{1}{6}\lambda_0 A^3\int_0^1 \sin^4(n\pi s)\,ds.$$

Using the integrals

$$\int_0^1 \sin^2(n\pi s)\,ds = \frac{1}{2}, \qquad \int_0^1 \sin^4(n\pi s)\,ds = \frac{3}{8},$$

we get

$$\lambda_2 A = \frac{1}{8}\lambda_0 A^3. \tag{3.30}$$

This is the bifurcation equation for the problem.

To rewrite (3.30) in terms of the original variables, let α denote the maximum value of a solution $\theta(s)$. Then, from (3.26) and (3.29), we have

$$\alpha = \varepsilon A + O\left(\varepsilon^3\right).$$

Using (3.30) in (3.25), we get the bifurcation equation

$$(\lambda - \lambda_0)\,\alpha = \frac{1}{8}\lambda_0\alpha^3 + O(\alpha^5) \qquad \text{as } \alpha \to 0. \tag{3.31}$$

Thus, in addition to the trivial solution $\alpha = 0$, we have solutions with

$$\alpha^2 = \frac{8\,(\lambda - \lambda_0)}{\lambda_0} + O\left(\alpha^4\right) \tag{3.32}$$

branching from each of the linearized eigenvalues λ_0 for $\lambda > \lambda_0$. This type of bifurcation is called a pitchfork bifurcation. It is supercritical because the new solutions appear for values of λ larger that the bifurcation value.

Thus, the original infinite-dimensional bifurcation problem (3.22)–(3.23) reduces to a one-dimensional bifurcation equation of the form $F(\alpha, \lambda) = 0$ in a neighborhood of the bifurcation point $(\theta, \lambda) = (0, \lambda_0)$. The bifurcation equation has the Taylor expansion (3.31) as $\alpha \to 0$ and $\lambda \to \lambda_0$.

6.2. Energy minimizers

For values of $\lambda > \pi^2$, solutions of the nonlinear BVP (3.22)–(3.23) are not unique. This poses the question of which solutions should be used. One criterion is that solutions of an equilibrium problem should be dynamically stable. We cannot address this question directly here, since we have not derived a set of time-dependent evolution equations. We can, however, use energy considerations.

The potential energy for (3.22) is

$$\mathcal{E}(\theta) = \int_0^1 \left\{\frac{1}{2}\left(\theta'\right)^2 - \lambda\left(1 - \cos\theta\right)\right\}ds. \tag{3.33}$$

We claim that the zero solution is a global minimizer of (3.33) when $\lambda \le \pi^2$, with $\mathcal{E}(0) = 0$, but it is not a minimizer when $\lambda > \pi$. As a result, the zero solution loses stability as λ passes through the first eigenvalue π^2, after which the rod will buckle.

To show that $\theta = 0$ is not a minimizer for $\lambda > \pi^2$, we compute the energy in the direction of the eigenvector of the first eigenvalue:

$$\begin{aligned}\mathcal{E}(\alpha \sin \pi s) &= \int_0^1 \left\{ \frac{1}{2}\alpha^2\pi^2 \cos^2 \pi s - \lambda \left[1 - \cos\left(\alpha \sin \pi s\right)\right] \right\} ds \\ &= \int_0^1 \left\{ \frac{1}{2}\alpha^2\pi^2 \cos^2 \pi s - \frac{1}{2}\alpha^2 \lambda \sin^2 \pi s \right\} ds + O\left(\alpha^4\right) \\ &= \frac{1}{4}\alpha^2 \left(\pi^2 - \lambda\right) + O\left(\alpha^4\right).\end{aligned}$$

It follows that we can have $\mathcal{E}(\theta) < \mathcal{E}(0)$ when $\lambda > \pi^2$.

For the converse, we use the Poincaré (or Wirtinger) inequality, which states that

$$\int_0^1 \theta^2 \, ds \le \frac{1}{\pi^2} \int_0^1 \theta'^2 \, ds$$

for all smooth functions such that $\theta(0) = 0$, $\theta(1) = 0$. (The inequality also holds for all $\theta \in H_0^1(0,1)$.) The best constant, $1/\pi^2$, in this inequality is the reciprocal of the lowest eigenvalue of (3.24), and it may be obtained by minimization of the corresponding Rayleigh quotient.

Using the inequality

$$1 - \cos\theta \le \frac{1}{2}\theta^2$$

in (3.33), followed by the Poincaré inequality, we see that

$$\mathcal{E}(\theta) \ge \int_0^1 \left\{ \frac{1}{2}\left(\theta'\right)^2 - \frac{1}{2}\theta^2 \right\} ds \ge \frac{1}{2}\left(1 - \frac{\lambda}{\pi^2}\right) \int_0^1 \left(\theta'\right)^2 \, ds.$$

It follows that $\mathcal{E}(\theta) \ge 0$ if $\lambda < \pi^2$, and $\theta = 0$ is the unique global minimizer of $\mathcal{E}$ among functions that vanish at the endpoints.

As the parameter λ passes through each eigenvalue $\lambda_0^{(n)}$, the energy function develops another direction (tangent to the corresponding eigenvector) in which it decreases as θ moves away from the critical point 0. These results are connected to conjugate points and Morse theory (see [**37**]).

The branches that bifurcate from $\lambda_0^{(n)}$ for $n \ge 2$ are of less interest than the first branch, because for $\lambda > \lambda_0^{(1)}$ we expect the solution to lie on one of the stable branches that bifurcates from $\lambda_0^{(1)}$ rather than on the trivial branch. We are then interested in secondary bifurcations of solutions from the stable branch rather than further bifurcations from the unstable trivial branch.

6.3. Solution by elliptic functions

Let us return to the solution of (3.22)–(3.23) in terms of elliptic functions.

The pendulum equation (3.22) has the first integral

$$\frac{1}{2}\left(\theta'\right)^2 + \lambda\left(1 - \cos\theta\right) = 2\lambda k^2$$

where k is a constant of integration; equivalently

$$\left(\theta'\right)^2 = 4\lambda\left(k^2 - \sin^2\frac{\theta}{2}\right).$$

Thus, if α is the maximum value of θ, we have

$$k = \sin\frac{\alpha}{2}. \tag{3.34}$$

Solving for θ', separating variables and integrating, we get

$$\int \frac{d\theta}{\sqrt{k^2 - \sin^2(\theta/2)}} = 2\sqrt{\lambda}s.$$

Here, the sign of the square root is chosen appropriately, and we neglect the constant of integration, which can be removed by a translation of s. Making the substitution $ku = \sin(\theta/2)$ in the integral, we get

$$\int \frac{du}{\sqrt{(1-u^2)\,(1-k^2u^2)}} = \sqrt{\lambda}s. \tag{3.35}$$

Trigonometric functions arise as inverse functions of integrals of the form

$$\int \frac{du}{\sqrt{p(u)}}$$

where $p(u)$ is a quadratic polynomial. In an analogous way, elliptic functions arise as inverse functions of integrals of the same form where $p(u)$ is a nondegenerate cubic or quartic polynomial. The Jacobi elliptic function $u \mapsto \mathrm{sn}(u,k)$, with modulus k, has the inverse function

$$\mathrm{sn}^{-1}(u,k) = \int_0^u \frac{dt}{\sqrt{(1-t^2)\,(1-k^2t^2)}}.$$

Rewriting u in terms of θ, it follows from (3.35) that $u = \mathrm{sn}(\sqrt{\lambda}s, k)$, so solutions $\theta(s)$ of (3.22) with $\theta(0) = 0$ are given by

$$\sin\left(\frac{\theta}{2}\right) = k\,\mathrm{sn}\left(\sqrt{\lambda}s, k\right).$$

The arclength ℓ of this solution from the endpoint $\theta = 0$ to the maximum deflection angle $\theta = \alpha$ is given by

$$\ell = \int_0^\ell ds = \int_0^\alpha \frac{d\theta}{\theta'}.$$

Using the substitution $ku = \sin(\theta/2)$, we get

$$\ell = \frac{1}{\sqrt{\lambda}}K(k)$$

where $K(k)$ is the complete elliptic integral of the first kind, defined by

$$K(k) = \int_0^1 \frac{du}{\sqrt{(1-u^2)\,(1-k^2u^2)}}. \tag{3.36}$$

This solution satisfies the boundary condition $\theta(1) = 0$ if $\ell = 1/(2n)$ for some integer $n = 1, 2, 3, \ldots$, meaning that

$$\lambda = 4n^2K^2(k). \tag{3.37}$$

This is the exact bifurcation equation for the n^{th} branch that bifurcates off the trivial solution.

A Taylor expansion of this equation agrees with the result from perturbation theory. From (3.36), we have, as $k \to 0$,

$$K(k) = \int_0^1 \frac{du}{\sqrt{1-u^2}} + \frac{1}{2}k^2 \int_0^1 \frac{u^2 du}{\sqrt{1-u^2}} + \cdots = \frac{\pi}{2}\left(1 + \frac{1}{4}k^2 + \dots\right).$$

Also, from (3.34), we have $k = \alpha/2 + \dots$. It follows that (3.37) has the expansion

$$\lambda = n^2\pi^2\left(1 + \frac{1}{4}k^2 + \dots\right)^2 = n^2\pi^2\left(1 + \frac{1}{8}\alpha^2 + \dots\right),$$

in agreement with (3.32).

There are also solutions with nonzero winding number, meaning that $\theta(0) = 0$ and $\theta(1) = 2\pi N$ for some nonzero integer N. These cannot be reached from the zero solution along a continuous branch, since the winding number is a discrete topological invariant.

6.4. Lyapounov-Schmidt reduction

The Lyapounov-Schmidt method provides a general approach to the rigorous derivation of local equilibrium bifurcation equations, based on an application of the implicit function theorem. We will outline the method and then explain how it applies to the buckling problem considered above. The main idea is to project the equation into two parts, one which can be solved uniquely and the other which gives the bifurcation equation.

Suppose that X, Y, Λ are Banach spaces, and $F : X \times \Lambda \to Y$ is a smooth map (at least C^1; see [**14**], for example, for more about derivatives of maps on Banach spaces). We are interested in solving the equation

$$F(x, \lambda) = 0 \tag{3.38}$$

for x as λ varies in the parameter space Λ.

We denote the partial derivatives of F at $(x, \lambda) \in X \times \Lambda$ by

$$F_x(x, \lambda) : X \to Y, \qquad F_\lambda(x, \lambda) : \Lambda \to Y.$$

These are bounded linear maps such that

$$F_x(x, \lambda)h = \frac{d}{d\varepsilon}F(x + \varepsilon h, \lambda)\bigg|_{\varepsilon=0}, \qquad F_\lambda(x, \lambda)\eta = \frac{d}{d\varepsilon}F(x, \lambda + \varepsilon\eta)\bigg|_{\varepsilon=0}.$$

Suppose that (x_0, λ_0) is a solution of (3.38), and denote by

$$L = F_x\left(x_0, \lambda_0\right) : X \to Y$$

the derivative of $F(x, \lambda)$ with respect to x at (x_0, λ_0).

The implicit function theorem states that if the bounded linear map L has an inverse $L^{-1} : Y \to X$, then (3.38) has a unique solution $x = f(\lambda)$ in some neighborhood of (x_0, λ_0). Moreover, the solution is at least as smooth as F, meaning that if F is C^k in a neighborhood of (x_0, λ_0), then f is C^k in a neighborhood of λ_0. Thus, roughly speaking, the nonlinear problem is locally uniquely solvable if the linearized problem is uniquely solvable.

It follows that a necessary condition for new solutions of (3.38) to bifurcate off a solution branch $x = f(\lambda)$ at (x_0, λ_0), where $x_0 = f\left(\lambda_0\right)$, is that $F_x\left(x_0, \lambda_0\right)$ is not invertible.

Consider such a point, and suppose that the non-invertible map $L : X \to Y$ is a Fredholm operator. This means that: (a) the null-space of L,

$$N = \{h \in X : Lh = 0\},$$

has finite dimension, and we can write $X = M \oplus N$ where M, N are closed subspaces of X; (b) the range of L,

$$R = \{k \in Y : k = Lh \text{ for some } h \in X\},$$

has finite codimension, and $Y = R \oplus S$ where R, S are closed subspaces of Y.

The condition that the range R of L is a closed subspace is satisfied automatically for maps on finite-dimensional spaces, but it is a significant assumption for maps on infinite-dimensional spaces. The condition that R has finite codimension simply means that any complementary space, such as S, has finite dimension (in which case the dimension does not depend on the choice of S).

We write $x \in X$ as $x = m + n$ where $m \in M$ and $n \in N$, and let

$$Q : Y \to Y$$

denote the projection onto R along S. That is, if $y = r + s$ is the unique decomposition of $y \in Y$ into a sum of $r \in R$ and $s \in S$, then $Qy = r$. Since R is closed, the linear map Q is bounded.

Equation (3.38) is equivalent to the pair of equations obtained by projecting it onto the range of L and the complementary space:

$$QF(m + n, \lambda) = 0, \tag{3.39}$$

$$(I - Q)F(m + n, \lambda) = 0. \tag{3.40}$$

We write (3.39) as

$$G(m, \nu) = 0,$$

where $\nu = (n, \lambda) \in \Gamma$, with $\Gamma = N \oplus \Lambda$, and $G : M \times \Gamma \to R$ is defined by

$$G(m, \nu) = QF(m + n, \lambda).$$

Let $x_0 = m_0 + n_0$ and $\nu_0 = (n_0, \lambda_0)$, so $(m_0, \nu_0) \in M \times \Gamma$ corresponds to $(x_0, \lambda_0) \in X \times \Lambda$. It follows from our definitions that the derivative of G

$$G_m(m_0, \nu_0) : M \to R$$

is an invertible linear map between Banach spaces. The implicit function theorem then implies that that (3.39) has a unique local solution for m of the form

$$m = g(n, \lambda)$$

where $g : N \times \Lambda \to M$.

Using this expression for m in (3.40), we find that (n, λ) satisfies an equation of the form

$$\Phi(n, \lambda) = 0 \tag{3.41}$$

where $\Phi : N \oplus \Lambda \to S$ is defined locally by

$$\Phi(n, \lambda) = (I - Q)F(g(n, \lambda) + n, \lambda).$$

Equation (3.41) is the bifurcation equation for (3.38). It describes all solutions of (3.38) in a neighborhood of a point (x_0, λ_0) where the derivative $F_x(x_0, \lambda_0)$ is singular.

This result is sometimes expressed in the following way. The m-component the solution $x = m + n$ is 'slaved' to the n-component; thus, if we can solve the

bifurcation equation for n in terms of λ, then m is determined by n. This allows us to reduce a larger bifurcation problem for $x \in X$ to a smaller bifurcation problem for $n \in N$.

If the null-space of L has dimension p and the range has codimension q, then (3.41) is equivalent to a system of p equations for q unknowns, depending on a parameter $\lambda \in \Lambda$. The integer $p - q$ is called the Fredholm index of L. In the commonly occuring case when the Fredholm index of L is zero, the bifurcation equation is a $p \times p$ system of equations. Thus, we can reduce bifurcation problems on infinite-dimensional spaces to ones on finite-dimensional spaces; the number of unknowns is equal to the dimension of the null space of L at the bifurcation point.

Next, we show how this method applies to the buckling problem. We write (3.22)–(3.23) as an equation $F(\theta, \lambda) = 0$, where $F : X \times \mathbb{R} \to Y$ is given by

$$F(\theta, \lambda) = \theta'' + \lambda \sin\theta$$

and

$$X = \left\{\theta \in H^2(0,1) : \theta(0) = 0,\, \theta(1) = 0\right\}, \qquad Y = L^2(0,1).$$

Here, $H^2(0,1)$ denotes the Sobolev space of functions whose weak derivatives of order less than or equal to 2 are square-integrable on $(0,1)$. Functions in $H^2(0,1)$ are continuously differentiable on $[0,1]$, so the boundary conditions make sense pointwise. Other function spaces, such as spaces of Hölder continuous functions, could be used equally well.

Consider bifurcations off the trivial solution $\theta = 0$. The derivative

$$L = F_\theta\left(0, \lambda_0\right)$$

is given by

$$Lh = h'' + \lambda_0 h.$$

This is singular on X if $\lambda_0 = n^2\pi^2$ for some $n \in \mathbb{N}$, so these are the only possible bifurcation points.

In this case, the null-space N of L is one-dimensional:

$$N = \left\{\alpha \sin(n\pi s) : \alpha \in \mathbb{R}\right\}.$$

We take as a closed complementary space

$$M = \left\{\varphi \in X : \textstyle\int_0^1 \varphi(s)\sin(n\pi s)\, ds = 0\right\}$$

The range R of L consists of the L^2-functions that are orthogonal to $\sin(n\pi s)$, meaning that

$$R = \left\{\rho \in L^2(0,1) : \textstyle\int_0^1 \rho(s)\sin(n\pi s)\, ds = 0\right\}.$$

As a complementary space, we take

$$S = \left\{\alpha \sin(n\pi s) : \alpha \in \mathbb{R}\right\}.$$

The projection $Q : L^2(0,1) \to L^2(0,1)$ onto R is then given by

$$(Q\rho)(s) = \rho(s) - \left[2\int_0^1 \rho\left(t\right)\sin\left(n\pi t\right)\, dt\right]\sin(n\pi s).$$

We write

$$\theta(s) = \varphi(s) + \alpha\sin(n\pi s)$$

where α is an arbitrary constant and $\varphi \in M$, so that

$$\int_0^1 \varphi(s) \sin(n\pi s)\, ds = 0. \tag{3.42}$$

In this case, equation (3.39) becomes

$$\begin{aligned} &\varphi'' + \lambda \sin\left[\varphi + \alpha \sin(n\pi s)\right] \\ &\qquad - 2\lambda \left\{ \int_0^1 \sin\left[\varphi\left(t\right) + \alpha \sin\left(n\pi t\right)\right] \sin\left(n\pi t\right)\, dt \right\} \sin(n\pi s) = 0, \end{aligned} \tag{3.43}$$

subject to the boundary conditions $\varphi(0) = 0$, $\varphi(1) = 0$, and the projection condition (3.42).

Equation (3.43) has the form $G(\varphi, \alpha, \lambda) = 0$, where $G : M \times \mathbb{R} \times \mathbb{R} \to R$. The derivative $G_\varphi\left(0, 0, \lambda_0\right) : M \to R$ is given by

$$G_\varphi\left(0, 0, \lambda_0\right) h(s) = h''(s) + \lambda_0 \left[h(s) - \left(2 \int_0^1 \sin\left(n\pi t\right) h(t)\, dt \right) \sin(n\pi s) \right].$$

It is one-to-one and onto, and has a bounded inverse. Therefore we can solve (3.43) locally for $\varphi(s) = g(s; a, \lambda)$. Equation (3.40) then gives the bifurcation equation

$$2\lambda \int_0^1 \sin\left[g(s; \alpha, \lambda) + \alpha \sin(n\pi s)\right] \sin(n\pi s)\, ds - \alpha\lambda_0 = 0. \tag{3.44}$$

A Taylor expansion of (3.43)–(3.44) in (α, λ) about $(0, \lambda_0)$ gives the same results as before.

Finally, we remark that these results, which are based on linearization and Taylor expansion, are local. There are also topological methods in bifurcation theory, introduced by Krasnoselski (1956) and Rabinowitz (1971), that use degree theory and provide global, but less explicit, results.

7. Laplace's equation

One of the most important variational principles for a PDE is Dirichlet's principle for the Laplace equation. We will show how Dirichlet's principle leads to the Laplace equation and describe how it arises in the potential theory of electrostatic fields.

7.1. Dirichlet principle

Let $\Omega \subset \mathbb{R}^n$ be a domain and $u : \overline{\Omega} \to \mathbb{R}$ a function. We assume that the domain and the function are sufficiently smooth.

The Dirichlet integral of u over Ω is defined by

$$\mathcal{F}(u) = \int_\Omega \frac{1}{2} \left|\nabla u\right|^2 dx. \tag{3.45}$$

Let us derive the Euler-Lagrange equation that must be satisfied by a minimizer of $\mathcal{F}$. To be specific, we consider a minimizer of $\mathcal{F}$ in a space of functions that satisfy Dirichlet conditions

$$u = f \qquad \text{on } \delta\Omega$$

where f is a given function defined on the boundary $\partial\Omega$ of Ω.

If $h : \overline{\Omega} \to \mathbb{R}$ is a function such that $h = 0$ on $\partial\Omega$, then

$$d\mathcal{F}(u)\, h = \left. \frac{d}{d\varepsilon} \int_\Omega \frac{1}{2} \left|\nabla u + \varepsilon \nabla h\right|^2 dx \right|_{\varepsilon = 0} = \int_\Omega \nabla u \cdot \nabla h\, dx.$$

Thus, any minimizer of the Dirichlet integral must satisfy

$$\int_\Omega \nabla u \cdot \nabla h \, dx = 0 \tag{3.46}$$

for all smooth functions h that vanish on the boundary.

Using the identity

$$\nabla \cdot (h \nabla u) = h \Delta u + \nabla u \cdot \nabla h$$

and the divergence theorem, we get

$$\int_\Omega \nabla u \cdot \nabla h \, dx = -\int_\Omega (\Delta u)\, h \, dx + \int_{\partial\Omega} h \frac{\partial u}{\partial n} \, dS.$$

Since $h = 0$ on $\partial\Omega$, the integral over the boundary is zero, and we get

$$d\mathcal{F}(u)\, h = -\int_\Omega (\Delta u)\, h \, dx$$

Thus, the variational derivative of $\mathcal{F}$, defined by

$$d\mathcal{F}(u)\, h = \int_\Omega \frac{\delta \mathcal{F}}{\delta u} h \, dx,$$

is given by

$$\frac{\delta \mathcal{F}}{\delta u} = -\Delta u.$$

Therefore, a smooth minimizer u of $\mathcal{F}$ satisfies Laplace's equation

$$\Delta u = 0. \tag{3.47}$$

This is the classical form of Laplace's equation, while (3.46) is the weak form.

Similarly, a minimizer of the functional

$$\mathcal{F}(u) = \int_\Omega \left\{ \frac{1}{2} |\nabla u|^2 - f u \right\} dx,$$

where $f : \overline{\Omega} \to \mathbb{R}$ is a given function, satisfies Poisson's equation

$$-\Delta u = f.$$

We will study the Laplace and Poisson equations in more detail later on.

7.2. The direct method

One of the simplest ways to prove the existence of solutions of the Laplace equation (3.47), subject, for example, to Dirichlet boundary conditions to show directly the existence of minimizers of the Dirichlet integral (3.45). We will not give any details here but we will make a few comments (see [**13**] for more information).

It was taken more-or-less taken for granted by Dirichlet, Gauss, and Riemann that since the Dirichlet functional (3.45) is a quadratic functional of u, which is bounded from below by zero, it attains its minimum for some function u, as would be the cases for such functions on $\mathbb{R}^n$. Weierstrass pointed out that this argument requires a nontrivial proof for functionals defined on infinite-dimensional spaces, because the Heine-Borel theorem that a bounded set is (strongly) precompact is not true in that case.

Let us give a few simple one-dimensional examples which illustrate the difficulties that can arise.

Example 3.18. Consider the functional (Weierstrass, 1895)

$$\mathcal{F}(u) = \frac{1}{2}\int_{-1}^{1} x^2 \left[u'(x)\right]^2 \, dx \tag{3.48}$$

defined on functions

$$u : [-1,1] \to \mathbb{R} \text{ such that } u(-1) = -1,\ u(1) = 1.$$

This functional is quadratic and bounded from below by zero. Furthermore, its infimum over smooth functions that satisfy the boundary conditions is equal to zero. To show this, for instance, let

$$u^{\varepsilon}(x) = \frac{\tan^{-1}(x/\varepsilon)}{\tan^{-1}(1/\varepsilon)} \qquad \text{for } \varepsilon > 0.$$

A straightforward computation gives

$$\mathcal{F}\left(u^{\varepsilon}\right) = \frac{\varepsilon}{\tan^{-1}(1/\varepsilon)} \to 0 \qquad \text{as } \varepsilon \to 0^{+}.$$

The Euler-Lagrange equation for (3.48) is

$$-\left[x^2 u'\right]' = 0.$$

Solutions u^{+}, u^{-} that satisfy the boundary conditions $u^{+}(1) = 1$, $u^{-}(-1) = -1$ have the form

$$u^{+}(x) = 1 + c^{+}\left(1 - \frac{1}{x}\right), \qquad u^{-}(x) = -1 + c^{-}\left(1 + \frac{1}{x}\right)$$

for some constants $c^{\pm}$. However, we cannot satisfy both boundary conditions for any choice the constants. Thus, there is no smooth, or even absolutely continuous, function that minimizes $\mathcal{F}$. Note that

$$\mathcal{F}(u) = \int_{-1}^{1} F\left(x, u'\right) \, dx, \qquad F(x,p) = \frac{1}{2}x^2 p^2.$$

The integrand $F(x,p)$ is a strictly convex function of p for each x, with

$$F_{pp}(x,p) = x^2 > 0,$$

except when $x = 0$. This loss of strict convexity at $x = 0$ is what leads to the singular behavior of the solutions of the Euler-Lagrange equations and the lack of a minimizer.

Example 3.19. Consider the functional

$$\mathcal{F}(u) = \int_{0}^{1} x^{2/3} \left[u'\right]^2 \, dx$$

defined on functions $u : [0,1] \to \mathbb{R}$ with $u(0) = 0$, $u(1) = 1$. The infimum is equal to zero. This infimum is attained for the function $u(x) = x^{1/3}$, which is not differentiable at $x = 0$. Thus, we cannot find a minimizer if we restrict the functional to C^1-functions; but we can find a minimizer on the larger class of absolutely continuous functions with weak derivative in $L^1(0,1)$. The minimizer is Hölder continuous with exponent $1/3$.

Example 3.20. Consider the non-convex functional

$$\mathcal{F}(u) = \int_0^1 \left(1 - [u']^2\right)^2 \, dx$$

defined on functions $u : [0,1] \to \mathbb{R}$ with $u(0) = 0$, $u(1) = 0$. The infimum is equal to zero. This infimum is not attained at any C^1-function, but it is attained at any 'zig-zag' Lipschitz continuous function that vanishes at the endpoints and whose derivative is equal to ± 1 almost everywhere. If we change the functional to

$$\mathcal{F}(u) = \int_0^1 \left\{u^2 + \left(1 - [u']^2\right)^2\right\} \, dx$$

then the infimum is still zero (as can be seen by taking a sequence of functions u_n with n 'zig-zags' and small L^∞-norm). This infimim, however, is not attained by any absolutely continuous function, since we cannot simultaneously make $|u'| = 1$ and $u = 0$. The difficulty here is associated with a lack of weak lower semicontinuity of the non-convex functional $\mathcal{F}$; for example, for the 'zig-zag' functions, we have $u_n \rightharpoonup 0$ in $W^{1,1}(0,1)$, but $\mathcal{F}(0) > \liminf_{n\to\infty} \mathcal{F}(u_n)$.

These difficulties were resolved for the Dirichlet functional by Hilbert (1900) and Lebesgue (1907), and Hilbert included several problems in the calculus of variations among his list of 23 problems at the 1900 ICM in Paris.

The Dirichlet functional is defined provided that ∇u is square-integrable. Thus, it is natural to look for minimizers of (3.45) is the Sobolev space $H^1(\Omega)$ of Lebesgue measurable, square-integrable functions $u : \Omega \to \mathbb{R}$ such that $u \in L^2(\Omega)$, meaning that $\int_\Omega u^2(x)\, dx < \infty$, with square-integrable weak derivatives $\partial_{x^i} u \in L^2(\Omega)$. If $g : \partial\Omega \to \mathbb{R}$ is a given boundary value that is attained by some function in $H^1(\Omega)$, then one can prove that there is a unique minimizer of (3.45) in the space

$$X = \left\{u \in H^1(\Omega) : \text{such that } u = g \text{ on } \partial\Omega\right\}.$$

The definition of the boundary values, or trace, of Sobolev functions requires a more careful discussion, but we will not go into the details here.

A further central issue in the calculus of variations is the regularity of minimizers. It is possible to prove that the minimizer of the Dirichlet functional is, in fact, a smooth function with continuous derivative of all orders inside Ω. In particular, it follows that it is a classical solution of Laplace's equation. Furthermore, if the boundary data and the domain are smooth, then the solution is also smooth on $\overline{\Omega}$.

7.3. Electrostatics

As an example of a physical problem leading to potential theory, consider a static electric field in a dielectric medium. (A dielectric medium is simply an insulator that does not conduct electricity, such as glass, air, or a vacuum.) We suppose that the dielectric has a charge-density $\rho(\vec{x})$, and that there is no magnetic field.

The electrostatic properties of the dielectric are characterized by two vector fields, the electric field $\vec{E}(\vec{x})$ and the electric displacement $\vec{D}(\vec{x})$. According to Maxwell's equations, these satisfy [**28**]

$$\operatorname{curl} \vec{E} = 0, \tag{3.49}$$

$$\operatorname{div} \vec{D} = \rho. \tag{3.50}$$

The integral form of these balance laws is

$$\int_\Gamma \vec{E} \cdot d\vec{x} = 0, \tag{3.51}$$

$$\int_{\partial\Omega} \vec{D} \cdot \vec{n}\, d\vec{x} = \int_\Omega \rho d\vec{x}, \tag{3.52}$$

for any closed curve Γ and any bounded volume Ω.

Equation (3.51) states that the circulation of $\vec{E}$ around the closed curve Γ is equal to zero, since by Stokes' theorem it is equal to the flux of $\operatorname{curl} \vec{E}$ through a surface bounded by Γ. Equation (3.52) states that the flux of $\vec{D}$ through a closed surface $\partial\Omega$ is equal to the total charge in the enclosed volume Ω.

On a simply connected domain, equation(3.49) implies that

$$\vec{E} = -\nabla\Phi.$$

for a suitable potential $\Phi(\vec{x})$.

The electric displacement is related to the electric field by a constitutive relation, which describes the response of the dielectric medium to an applied electric field. We will assume that it has the simplest linear, isotropic form

$$\vec{E} = \epsilon\vec{D}$$

where ϵ is a constant, called the dielectric constant, or electric permittivity, of the medium. In a linear, anisotropic medium, ϵ becomes a tensor; for large electric fields, it may be necessary to use a nonlinear constitutive relation.

It follows from these equations and (3.50) that Φ satisfies Poisson's equation

$$-\epsilon\Delta\Phi = \rho. \tag{3.53}$$

This equation is supplemented by boundary conditions; for example, we require that Φ is constant on a conducting boundary, and the normal derivative of Φ is zero on an insulating boundary.

The energy of the electrostatic field in some region $\Omega \subset \mathbb{R}^3$ is

$$\mathcal{E} = \int_\Omega \left\{ \frac{1}{2}\vec{E} \cdot \vec{D} - \rho\Phi \right\} d\vec{x} = \int_\Omega \left\{ \frac{1}{2}\epsilon \left|\nabla\Phi\right|^2 - \rho\Phi \right\} d\vec{x}.$$

The term proportional to $\vec{E} \cdot \vec{D}$ the the energy of the field, while the term proportional to $\rho\Phi$ is the work required to bring the charge distribution to the potential Φ.

The potential Φ minimizes this functional, and the condition that $\mathcal{E}(\Phi)$ is stationary with respect to variations in Φ leads to (3.53).

8. The Euler-Lagrange equation

A similar derivation of the Euler-Lagrange equation as the condition satisfied by a smooth stationary point applies to more general functionals. For example, consider the functional

$$\mathcal{F}(u) = \int_\Omega F\left(x, u, \nabla u\right) dx$$

where $\Omega \subset \mathbb{R}^n$ and $u : \overline{\Omega} \to \mathbb{R}^m$. Then, writing

$$x = \left(x^1, x^2, \ldots, x^n\right), \qquad u = \left(u^1, u^2, \ldots, u^m\right),$$

denoting a derivative with respect to x^j by ∂_j, and using the summation convention, we have

$$d\mathcal{F}(u)h = \int_\Omega \left\{ F_{u^i}\left(x, u, \nabla u\right) h^i + F_{\partial_j u^i}\left(x, u, \nabla u\right) \partial_j h^i \right\} \, dx.$$

Thus, u is a stationary point of $\mathcal{F}$ if

$$\int_\Omega \left\{ F_{u^i}\left(x, u, \nabla u\right) h^i + F_{\partial_j u^i}\left(x, u, \nabla u\right) \partial_j h^i \right\} \, dx = 0 \tag{3.54}$$

for all smooth test functions $h : \overline{\Omega} \to \mathbb{R}$ that vanish on the boundary.

Using the divergence theorem, we find that

$$d\mathcal{F}(u)h = \int_\Omega \left\{ F_{u^i}\left(x, u, \nabla u\right) - \partial_j \left[F_{\partial_j u^i}\left(x, u, \nabla u\right) \right] \right\} h^i \, dx.$$

Thus,

$$\frac{\delta \mathcal{F}}{\delta h^i} = -\partial_j \left[F_{\partial_j u^i}\left(x, u, \nabla u\right) \right] + F_{u^i}\left(x, u, \nabla u\right),$$

and a smooth stationary point u satisfies

$$-\partial_j \left[F_{\partial_j u^i}\left(x, u, \nabla u\right) \right] + F_{u^i}\left(x, u, \nabla u\right) = 0 \qquad \text{for } i = 1, 2, \ldots, n.$$

The weak form of this equation is (3.54).

8.1. The minimal surface equation

Suppose that a surface over a domain $\Omega \subset \mathbb{R}^n$ is the graph of a smooth function $z = u(x)$, where $u : \overline{\Omega} \to \mathbb{R}$. The area $\mathcal{A}$ of the surface is

$$\mathcal{A}(u) = \int_\Omega \sqrt{1 + |\nabla u|^2} \, dx.$$

The problem of finding a surface of minimal area that spans a given curve $z = g(x)$ over the boundary, where $g : \partial\Omega \to \mathbb{R}$, is called Plateau's problem. Any smooth minimizer of the area functional $\mathcal{A}(u)$ must satisfy the Euler-Lagrange equation, called the minimal surface problem,

$$\nabla \cdot \left[\frac{\nabla u}{\sqrt{1 + |\nabla u|^2}} \right] = 0.$$

As a physical example, a film of soap has energy per unit area equal to its surface tension. Thus, a soap file on a wire frame is a minimal surface.

A full analysis of this problem is not easy. The PDE is elliptic, but it is nonlinear and it is not uniformly elliptic, and it has motivated a large amount of work on quasilinear elliptic PDEs. See [**13**] for more information.

8.2. Nonlinear elasticity

Consider an equilibrium deformation of an elastic body. We label material points by their location $\vec{x} \in \mathcal{B}$ in a suitable reference configuration $\mathcal{B} \subset \mathbb{R}^n$. A deformation is described by an invertible function $\vec{\varphi} : \mathcal{B} \to \mathbb{R}^n$, where $\vec{\varphi}\,(\vec{x})$ is the location of the material point $\vec{x}$ in the deformed configuration of the body.

The deformation gradient

$$\mathbf{F} = \nabla \vec{\varphi}, \qquad F_{ij} = \frac{\partial \varphi_i}{\partial x_j}$$

gives a linearized approximation of the deformation at each point, and therefore describes the local strain and rotation of the deformation.

An elastic material is said to be hyperelastic if the work required to deform it, per unit volume in the reference configuration, is given by a scalar-valued strain energy function. Assuming, for simplicity, that the body is homogeneous so the work does not depend explicitly on $\vec{x}$, the strain energy is a real-valued function $W(\mathbf{F})$ of the deformation gradient. In the absence of external forces, the total energy of a deformation $\vec{\varphi}$ is given by

$$\mathcal{W}(\vec{\varphi}) = \int_{\mathcal{B}} W(\nabla\vec{\varphi}(\vec{x}))\, d\vec{x}.$$

Equilibrium deformations are minimizers of the total energy, subject to suitable boundary conditions. Therefore, smooth minimizers satisfy the Euler-Lagrange equations

$$\nabla \cdot \mathbf{S}(\nabla\vec{\varphi}(\vec{x})) = 0, \qquad \frac{\partial S_{ij}}{\partial x_j} = 0 \quad i = 1, \dots, n,$$

where we use the summation convention in the component form of the equations, and $\mathbf{S}(\mathbf{F})$ is the Piola-Kirchoff stress tensor, given by

$$\mathbf{S} = \nabla_{\mathbf{F}} W, \qquad S_{ij} = \frac{\partial W}{\partial F_{ij}}.$$

This is an $n \times n$ nonlinear system of second-order PDEs for $\vec{\varphi}(\vec{x})$.

There are restrictions on how the strain energy W depends on $\mathbf{F}$. The principle of material frame indifference [**25**] implies that, for any material,

$$W(\mathbf{RF}) = W(\mathbf{F})$$

for all orthogonal transformations $\mathbf{R}$. If the elastic material is isotropic, then

$$W(\mathbf{F}) = \tilde{W}(\mathbf{B})$$

depends only on the left Cauchy-Green strain tensor $\mathbf{B} = \mathbf{F}\mathbf{F}^{\top}$, and, in fact, only on the principle invariants of $\mathbf{B}$.

The constitutive restriction imply that W is not a convex function of $\mathbf{F}$. This creates a difficulty in the proof of the existence of minimizers by the use of direct methods, because one cannot use convexity to show that $\mathcal{W}$ is weakly lower semicontinuous.

This difficult was overcome by Ball (1977). He observed that one can write

$$W(\mathbf{F}) = \hat{W}(\mathbf{F}, \operatorname{cof}\mathbf{F}, \det\mathbf{F})$$

where $\hat{W}$ is a convex function of $\mathbf{F}$ and the cofactors $\operatorname{cof}\mathbf{F}$ of $\mathbf{F}$, including its determinant. A function with this property is said to be *polyconvex*.

According to the theory of compensated compactness, given suitable bounds on the derivatives of $\vec{\varphi}$, the cofactors of $\mathbf{F}$ are weakly continuous, which is a very unusual property for nonlinear functions. Using this fact, combined with the observation that the strain energy is polyconvex, Ball was able to prove the existence of minimizers for nonlinear hyperelasticity.

9. The wave equation

Consider the motion of a medium whose displacement may be described by a scalar function $u(x,t)$, where $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$. For example, this function might represent the transverse displacement of a membrane $z = u(x,y,t)$.

Suppose that the kinetic energy $\mathcal{T}$ and potential energy $\mathcal{V}$ of the medium are given by

$$\mathcal{T}(u_t) = \frac{1}{2}\int \rho_0 u_t^2\, dx, \qquad \mathcal{V}(u) = \frac{1}{2}\int k\,|\nabla u|^2\, dx,$$

where $\rho_0(\vec{x})$ is a mass-density and $k(\vec{x})$ is a stiffness, both assumed positive. The Lagrangian $\mathcal{L} = \mathcal{T} - \mathcal{V}$ is

$$\mathcal{L}(u,u_t) = \int \frac{1}{2}\left\{\rho_0 u_t^2 - k\,|\nabla u|^2\right\} dx, \tag{3.55}$$

and the action — the time integral of the Lagrangian — is

$$\mathcal{S}(u) = \int\int \frac{1}{2}\left\{\rho_0 u_t^2 - k\,|\nabla u|^2\right\} dxdt.$$

Note that the kinetic and potential energies and the Lagrangian are functionals of the spatial field and velocity $u(\cdot,t)$, $u_t(\cdot,t)$ at each fixed time, whereas the action is a functional of the space-time field $u(x,t)$, obtained by integrating the Lagrangian with respect to time.

The Euler-Lagrange equation satisfied by a stationary point of this action is

$$\rho_0 u_{tt} - \nabla\cdot(k\nabla u) = 0. \tag{3.56}$$

If ρ_0, k are constants, then

$$u_{tt} - c_0^2 \Delta u = 0, \tag{3.57}$$

where $c_0^2 = k/\rho_0$. This is the linear wave equation with wave-speed c_0.

Unlike the energy for Laplace's equation, the action functional for the wave equation is not positive definite. We therefore cannot expect a solution of the wave equation to be a minimizer of the action, in general, only a critical point. As a result, direct methods are harder to implement for the wave equation (and other hyperbolic PDEs) than they are for Laplace's equation (and other elliptic PDEs), although there are 'mountain-pass' lemmas that can be used to establish the existence of stationary points. Moreover, in general, stationary points of functionals do not have the increased regularity that minimizers of convex functionals typically possess.

10. Hamiltonian mechanics

Let us return to the motion of a particle in a conservative force field considered in Section 1. We will give an alternative, Hamiltonian, formulation of its equations of motion.

Given a Lagrangian

$$L(\vec{x},\vec{v}) = \frac{1}{2}m\,|\vec{v}|^2 - V(\vec{x}),$$

we define the momentum $\vec{p}$ by

$$\vec{p} = \frac{\partial L}{\partial \vec{v}}, \tag{3.58}$$

meaning that $\vec{p} = m\vec{v}$. Here, we use the notation

$$\frac{\partial L}{\partial \vec{v}} = \left(\frac{\partial L}{\partial v_1}, \frac{\partial L}{\partial v_2}, \ldots, \frac{\partial L}{\partial v_n} \right)$$

to denote the derivative with respect to $\vec{v}$, keeping $\vec{x}$ fixed, with a similar notation for the derivative $\partial/\partial \vec{p}$ with respect to $\vec{p}$, keeping $\vec{x}$ fixed. The derivative $\partial/\partial \vec{x}$ is taken keeping $\vec{v}$ or $\vec{p}$ fixed, as appropriate.

We then define the Hamiltonian function H by

$$H(\vec{x}, \vec{p}) = \vec{p} \cdot \vec{v} - L(\vec{x}, \vec{v}), \tag{3.59}$$

where we express $\vec{v} = \vec{p}/m$ on the right hand side in terms of $\vec{p}$. This gives

$$H(\vec{x}, \vec{p}) = \frac{1}{2m} \vec{p} \cdot \vec{p} + V(\vec{x}).$$

Thus, we transform L as a function of $\vec{v}$ into H as a function of $\vec{p}$. The variable $\vec{x}$ plays the role of a parameter in this transformation. The function $H(\vec{x}, \vec{p})$, given by (3.58)–(3.59) is the Legendre transform of $L(\vec{x}, \vec{v})$ with respect to $\vec{v}$; conversely, L is the Legendre transform of H with respect to $\vec{p}$.

Note that the the Hamiltonian is the total energy of the particle,

$$H(\vec{x}, \vec{p}) = T(\vec{p}) + V(\vec{x}),$$

where T is the kinetic energy expressed as a function of the momentum

$$T(\vec{p}) = \frac{1}{2m} |\vec{p}|^2.$$

The Lagrangian equation of motion (3.1) may then be written as a first order system for $(\vec{x}, \vec{p})$:

$$\dot{\vec{x}} = \frac{1}{m} \vec{p}, \qquad \dot{\vec{p}} = -\frac{\partial V}{\partial \vec{x}}.$$

This system has the canonical Hamiltonian form

$$\dot{\vec{x}} = \frac{\partial H}{\partial \vec{p}}, \qquad \dot{\vec{p}} = -\frac{\partial H}{\partial \vec{x}}. \tag{3.60}$$

Equation (3.60) is a $2n$-dimensional system of first-order equations. We refer to the space $\mathbb{R}^{2n}$, with coordinates $(\vec{x}, \vec{p})$, as the phase space of the system.

10.1. The Legendre transform

The above transformation, from the Lagrangian as a function of velocity to the Hamiltonian as a function of momentum, is an example of a Legendre transform. In that case, the functions involved were quadratic.

More generally, if $f : \mathbb{R}^n \to \mathbb{R}$, we define the Legendre transform $f^*(x^*)$ of the function $f(x)$ as follows. Let

$$x^* = \frac{\partial f}{\partial x}(x),$$

and suppose we can invert this equation to get $x = x(x^*)$. This is the case, for example, if f is a smooth, convex function. We then define

$$f^*(x^*) = x^* \cdot x(x^*) - f(x(x^*)).$$

Note that, by the chain rule and the definition of x^*, we have

$$\frac{\partial f^*}{\partial x^*} = x + x^* \cdot \frac{\partial x}{\partial x^*} - \frac{\partial f}{\partial x} \frac{\partial x}{\partial x^*} = x + x^* \cdot \frac{\partial x}{\partial x^*} - x^* \cdot \frac{\partial x}{\partial x^*} = x,$$

and, from the definition of f^*,

$$f(x) = x \cdot x^*(x) - f^*(x^*(x)).$$

Thus, if f is convex, the Legendre transform of $f^*(x^*)$ is the original function $f(x)$ (see [**43**] for more on convex analysis).

Consider a Lagrangian $F(x, u, u')$, where $u : [a, b] \to \mathbb{R}^n$ and $F : [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. Taking the Legendre transform of F with respect to the u'-variable, we get

$$p = \frac{\partial F}{\partial u'}, \qquad H(x, u, p) = p \cdot u' - F(x, u, u').$$

It follows that

$$\frac{\partial H}{\partial u} = p \cdot \frac{\partial u'}{\partial u} - \frac{\partial F}{\partial u} - \frac{\partial F}{\partial u'}\frac{\partial u'}{\partial u} = -\frac{\partial F}{\partial u},$$
$$\frac{\partial H}{\partial p} = u' + p \cdot \frac{\partial u'}{\partial p} - \frac{\partial F}{\partial u'}\frac{\partial u'}{\partial p} = u'.$$

Hence, the Euler-Lagrange equation

$$-\frac{d}{dx}\left(\frac{\partial F}{\partial u'}\right) + \frac{\partial F}{\partial u} = 0$$

may be written as a Hamiltonian system

$$u' = \frac{\partial H}{\partial p}, \qquad p' = -\frac{\partial H}{\partial u}.$$

In general, the Hamiltonian in these equations may depend on the independent variable x (or t in the mechanical problem above) as well as the dependent variables. For simplicity, we will consider below Hamiltonians that do not depend explicitly on the independent variable.

10.2. Canonical coordinates

It is important to recognize that there are two ingredients in the Hamiltonian system (3.60). One is obvious: the Hamiltonian function $H(\vec{x}, \vec{p})$ itself. The other, less obvious, ingredient is a Hamiltonian structure that allows us to map the differential of a Hamiltonian function

$$dH = \frac{\partial H}{\partial \vec{x}}\, d\vec{x} + \frac{\partial H}{\partial \vec{p}}\, d\vec{p}$$

to the Hamiltonian vector field

$$X_H = \frac{\partial H}{\partial \vec{p}}\,\frac{\partial}{\partial \vec{x}} - \frac{\partial H}{\partial \vec{x}}\,\frac{\partial}{\partial \vec{p}}$$

that appears in (3.60).

We will not describe the symplectic geometry of Hamiltonian systems in any detail here (see [**6**] for more information, including an introduction to differential forms) but we will make a few comments to explain the role of canonical coordinates $(\vec{x}, \vec{p})$ in the formulation of Hamiltonian equations.

The Hamitonian structure of (3.60) is defined by a symplectic two-form

$$\omega = d\vec{x} \wedge d\vec{p} \tag{3.61}$$

on the phase space $\mathbb{R}^{2n}$. More generally, one can consider symplectic manifolds, which are manifolds, necessarily even-dimensional, equipped with a closed, nondegenerate two-form ω.

The two-form (3.61) can be integrated over a two-dimensional submanifold S of $\mathbb{R}^{2n}$ to give an 'area'

$$\int_S \omega = \sum_{i=1}^{n} \int_S dx^i \wedge dp_i.$$

Roughly speaking, this integral is the sum of the oriented areas of the projections of S, counted according to multiplicity, onto the $\left(x^i, p_i\right)$-coordinate planes. Thus, the phase space of a Hamiltonian system has a notion of oriented area, defined by the *skew-symmetric* two-form ω. In a somewhat analogous way, Euclidean space (or a Riemannian manifold) has a notion of length and angle, which is defined by a *symmetric* two-form, the metric g. The geometry of symplectic manifolds (M, ω) is, however, completely different from the more familiar geometry of Riemannian manifolds (M, g).

According to Darboux's theorem, if ω is a closed nondegenerate two-form, then there are local coordinates $(\vec{x}, \vec{p})$ in which it is given by (3.61). Such coordinates are called canonical coordinates, and Hamilton's equations take the canonical form (3.60) for every Hamiltonian H in any canonical system of coordinates. The canonical form of ω and Hamilton's equations, however, are not preserved under arbitrary transformations of the dependent variables.

A significant part of the theory of Hamiltonian systems, such as Hamilton-Jacobi theory, is concerned with finding canonical transformations that simplify Hamilton's equations. For example, if, for a given Hamiltonian $H\left(\vec{x}, \vec{p}\right)$, we can find a canonical change of coordinates such that

$$\left(\vec{x}, \vec{p}\right) \mapsto \left(\vec{x}', \vec{p}'\right), \qquad H\left(\vec{x}, \vec{p}\right) \mapsto H\left(\vec{p}'\right),$$

meaning that the transformed Hamiltonian is independent of the position variable $\vec{x}'$, then we can solve the corresponding Hamiltonian equations explicitly. It is typically not possible to do this, but the completely integrable Hamiltonian systems for which it is possible form an important and interesting class of solvable equations. We will not discuss these ideas further here (see [**24**] for more information).

11. Poisson brackets

It can be inconvenient to use conjugate variables, and in some problems it may be difficult to identify which variables form conjugate pairs. The Poisson bracket provides a way to write Hamiltonian systems, as well as odd-order generalizations of the even-order canonical systems, which does not require the use of canonical variables. The Poisson bracket formulation is also particularly convenient for the description of Hamiltonian PDEs.

First we describe the Poisson-bracket formulation of the canonical equations. Let $\vec{u} = \left(\vec{x}, \vec{p}\right)^\top \in \mathbb{R}^{2n}$. Then we may write (3.60) as

$$\dot{\vec{u}} = \mathbf{J} \frac{\partial H}{\partial \vec{u}}$$

where $\mathbf{J} : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ is the constant skew-symmetric linear map with matrix

$$\mathbf{J} = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}. \tag{3.62}$$

If $F, G : \mathbb{R}^{2n} \to \mathbb{R}$ are smooth functions, then we define their Poisson bracket $\{F, G\}$, which is also a function $\{F, G\} : \mathbb{R}^{2n} \to \mathbb{R}$, by

$$\{F, G\} = \frac{\partial F}{\partial \vec{u}} \cdot \mathbf{J} \frac{\partial G}{\partial \vec{u}}.$$

In terms of derivatives with respect to $(\vec{x}, \vec{p})$, the bracket is given by

$$\{F, G\} = \frac{\partial F}{\partial \vec{x}} \cdot \frac{\partial G}{\partial \vec{p}} - \frac{\partial F}{\partial \vec{p}} \cdot \frac{\partial G}{\partial \vec{x}} = \sum_{i=1}^{n} \left(\frac{\partial F}{\partial x^i} \frac{\partial G}{\partial p_i} - \frac{\partial F}{\partial p_i} \frac{\partial G}{\partial x^i} \right) \tag{3.63}$$

Hamilton's equations may be written as

$$\dot{\vec{u}} = \{\vec{u}, H\},$$

or, in component form,

$$u^i = \left\{u^i, H\right\} \qquad 1 \leq i \leq 2n.$$

Moreover, if $F(\vec{u})$ is any function, then

$$\dot{F} = \frac{\partial F}{\partial \vec{x}} \dot{\vec{x}} + \frac{\partial F}{\partial \vec{p}} \dot{\vec{p}} = \frac{\partial F}{\partial \vec{x}} \frac{\partial H}{\partial \vec{p}} - \frac{\partial F}{\partial \vec{p}} \frac{\partial H}{\partial \vec{x}}.$$

It follows that

$$\dot{F} = \{F, H\}.$$

Thus, a function $F(\vec{x}, \vec{p})$ that does not depend explicitly on time t is a conserved quantity for Hamilton's equations if its Poisson bracket with the Hamiltonian vanishes; for example, the Poisson bracket of the Hamiltonian with itself vanishes, so the Hamiltonian is conserved.

The Poisson bracket in (3.63) has the properties that for any functions F, G, H and constants a, b

$$\{F, G\} = -\{G, F\}, \tag{3.64}$$

$$\{aF + bG, H\} = a\{F, H\} + b\{G, H\}, \tag{3.65}$$

$$\{FG, H\} = F\{G, H\} + \{F, H\}G, \tag{3.66}$$

$$\{F, \{G, H\}\} + \{G, \{H, F\}\} + \{H, \{F, G\}\} = 0. \tag{3.67}$$

That is, it is skew-symmetric (3.64), bilinear (3.65), a derivation (3.66), and satisfies the Jacobi identity (3.67).

Any bracket with these properties that maps a pair of smooth functions F, G to a smooth function $\{F, G\}$ defines a Poisson structure. The bracket corresponding to the matrix $\mathbf{J}$ in (3.62) is the canonical bracket, but there are many other brackets. In particular, the skew-symmetric linear operator $\mathbf{J}$ can depend on u, provided that the associated bracket satisfies the Jacobi identity.

12. Rigid body rotations

Consider a rigid body, such as a satellite, rotating about its center of mass in three space dimensions.

We label the material points of the body by their position $\vec{a} \in \mathcal{B}$ in a given reference configuration $\mathcal{B} \subset \mathbb{R}^3$, and denote the mass-density of the body by

$$\rho : \mathcal{B} \to [0, \infty).$$

We use coordinates such that the center of mass of the body at the origin, so that

$$\int_{\mathcal{B}} \rho\,(\vec{a})\,\vec{a}\,d\vec{a} = 0.$$

Here, $d\vec{a}$ denotes integration with respect to volume in the reference configuration.

The possible configurations of the body are rotations of the reference configuration, so the configuration space of a the body may be identified with the rotation group. This is the special orthogonal group $SO(3)$ of linear transformations R on $\mathbb{R}^3$ such that

$$R^\top R = I, \qquad \det R = 1.$$

The first condition is the orthogonality condition, $R^\top = R^{-1}$, which ensures that R preserves the Euclidean inner product of vectors, and therefore lengths and angles. The second condition restricts R to the 'special' transformations with determinant one. It rules out the orientation-reversing orthogonal transformations with $\det R = -1$, which are obtained by composing a reflection and a rotation.

First, we will define the angular velocity and angular momentum of the body in a spatial reference frame. Then we will 'pull back' these vectors to the body reference frame, in which the equations of motion simplify.

12.1. Spatial description

Consider the motion of the body in an inertial frame of reference whose origin is at the center of mass of the body. The position vector $\vec{x}$ of a point $\vec{a} \in \mathcal{B}$ at time t is given by

$$\vec{x}\,(\vec{a}, t) = R(t)\vec{a} \tag{3.68}$$

where $R(t) \in SO(3)$ is a rotation. Thus, the motion of the rigid body is described by a curve of rotations $R(t)$ in the configuration space $SO(3)$.

Differentiating (3.68) with respect to t, and using (3.68) in the result, we find that the velocity

$$\vec{v}\,(\vec{a}, t) = \dot{\vec{x}}\,(\vec{a}, t)$$

of the point $\vec{a}$ is given by

$$\vec{v} = w\vec{x}, \tag{3.69}$$

where

$$w = \dot{R}R^\top. \tag{3.70}$$

Differentiation of the equation $RR^\top = I$ with respect to t implies that

$$\dot{R}R^\top + R\dot{R}^\top = 0.$$

Thus, w in (3.70) is skew-symmetric, meaning that $w^\top = -w$.

If $W : \mathbb{R}^3 \to \mathbb{R}^3$ is a skew-symmetric linear map on three-dimensional Euclidean space, then there is a unique vector $\vec{\Omega} \in \mathbb{R}^3$ such that

$$W\vec{x} = \vec{\Omega} \times \vec{x}. \tag{3.71}$$

We denote this correspondence by $\vec{\Omega} = \hat{W}$. With respect to a right-handed orthonormal basis, the matrix of W and the components of $\vec{\Omega}$ are related by

$$\begin{pmatrix} 0 & -\Omega_3 & \Omega_2 \\ \Omega_3 & 0 & -\Omega_1 \\ -\Omega_2 & \Omega_1 & 0 \end{pmatrix} \longleftrightarrow \begin{pmatrix} \Omega_1 \\ \Omega_2 \\ \Omega_3 \end{pmatrix}$$

We let $\vec{\omega}(t) = \hat{w}(t)$ denote the vector associated with w in (3.70). Then, from (3.69), the velocity of the body in the spatial frame is

$$\vec{v} = \vec{\omega} \times \vec{x}. \tag{3.72}$$

Thus, the vector $\vec{\omega}(t)$ is the angular velocity of the body at time t.

The angular momentum π, or moment of momentum, of the body is defined by

$$\pi(t) = \int_{\mathcal{B}} \rho(\vec{a}) \left[\vec{x}(\vec{a}, t) \times \vec{v}(\vec{a}, t)\right] d\vec{a} \tag{3.73}$$

Equivalently, making the change of variables $\vec{a} \mapsto \vec{x}$ in (3.68) in the integral, whose Jacobian is equal to one, we get

$$\pi = \int_{\mathcal{B}_t} \rho\left(R^\top(t)\vec{a}\right) \left[\vec{x} \times (\vec{\omega}(t) \times \vec{x})\right] d\vec{x},$$

where $\mathcal{B}_t = \vec{x}(\mathcal{B}, t)$ denotes the region occupied by the body at time t.

Conservation of angular momentum implies that, in the absence of external forces and couples,

$$\dot{\vec{\pi}} = 0. \tag{3.74}$$

This equation is not so convenient to solve for the motion of the body, because the angular momentum $\vec{\pi}$ depends in a somewhat complicated way on the angular velocity $\vec{\omega}$ and the rotation matrix R. We will rewrite it with respect to quantities defined with respect to the body frame, which leads to a system of ODEs for the angular momentum, or angular velocity, in the body frame.

12.2. Body description

The spatial coordinate $\vec{x}$ is related to the body coordinate $\vec{a}$ by $\vec{x} = R\vec{a}$. Similarly, we define a body frame velocity $\vec{V}(\vec{a}, t)$, angular velocity $\Omega(t)$, and angular momentum $\Pi(t)$ in terms of the corresponding spatial vectors by

$$\vec{v} = R\vec{V}, \qquad \vec{\omega} = R\vec{\Omega}, \qquad \vec{\pi} = R\vec{\Pi}. \tag{3.75}$$

Thus, we rotate the spatial vectors back to the body frame.

First, from (3.69), we find that if $\vec{v} = R\vec{V}$, then

$$\vec{V} = W\vec{a} \tag{3.76}$$

where W is the skew-symmetric map

$$W = R^\top \dot{R}. \tag{3.77}$$

Therefore, denoting by $\vec{\Omega} = \hat{W}$ the vector associated with W, we have

$$\vec{V} = \vec{\Omega} \times \vec{a}. \tag{3.78}$$

Since $w = RWR^\top$, it follows that $\vec{\omega} = R\vec{\Omega}$, as in (3.75).

Next, since rotations preserve cross-products, we have

$$\vec{x} \times \vec{v} = R\left(\vec{a} \times \vec{V}\right).$$

Using this equation, followed by (3.78), in (3.73), we find that $\vec{\pi} = R\vec{\Pi}$ where

$$\vec{\Pi}(t) = \int_{\mathcal{B}} \rho(\vec{a}) \left[\vec{a} \times (\Omega(t) \times \vec{a})\right] d\vec{a}.$$

This equation is a linear relation between the angular velocity and angular momentum. We write it as

$$\vec{\Pi} = \mathbf{I}\vec{\Omega}. \tag{3.79}$$

where $\mathbf{I} : \mathbb{R}^3 \to \mathbb{R}^3$ is a constant linear map depending only on the mass distribution of the body. It is called the *inertia tensor*.

An explicit expression for the inertia tensor is

$$\mathbf{I} = \int_{\mathcal{B}} \rho\left(\vec{a}\right) \left[\left(\vec{a} \cdot \vec{a}\right) I - \vec{a} \otimes \vec{a}\right] \, d\vec{a}, \tag{3.80}$$

where I denotes the identity transformation, or, in components,

$$I_{ij} = \int_{\mathcal{B}} \rho\left(\vec{a}\right) \left[\left(a_k a_k\right) \delta_{ij} - a_i a_j\right] \, d\vec{a}.$$

The inertia tensor is symmetric and positive definite. In the limiting case of a rod, or 'rotator,' idealized as a straight line with a mass density per unit length, the eigenvalue of $\mathbf{I}$ corresponding to rotations about the axis of the rod is zero, and $\mathbf{I}$ is singular. We will not consider that case here, and assume that $\mathbf{I}$ is nonsingular.

The quantities in (3.79) have dimensions

$$\left[\vec{\Omega}\right] = \frac{1}{T}, \qquad \left[\vec{\Pi}\right] = \frac{ML^2}{T}, \qquad [\mathbf{I}] = ML^2,$$

so the equation is dimensionally consistent.

Using the equation $\vec{\pi} = R\vec{\Pi}$ in the spatial equation of conservation of angular momentum (3.74), using (3.77) to write $\dot{R}$ in terms of W, and using the fact that $\vec{\Omega} = \hat{W}$, we get the body form of conservation of angular momentum

$$\dot{\vec{\Pi}} + \Omega \times \vec{\Pi} = 0.$$

Together with (3.79), this equation provides a 3×3 system of ODEs for either the body angular velocity $\vec{\Omega}(t)$

$$\mathbf{I}\dot{\vec{\Omega}} + \vec{\Omega} \times \left(\mathbf{I}\vec{\Omega}\right) = 0,$$

or the body angular momentum $\vec{\Pi}(t)$

$$\dot{\vec{\Pi}} + \left(\mathbf{I}^{-1}\vec{\Pi}\right) \times \vec{\Pi} = 0. \tag{3.81}$$

Once we have solved these equations for $\vec{\Omega}(t)$, and therefore $W(t)$, we may reconstruct the rotation $R(t)$ by solving the matrix equation

$$\dot{R} = RW.$$

12.3. The kinetic energy

The kinetic energy T of the body is given by

$$T = \frac{1}{2} \int_{\mathcal{B}} \rho\left(\vec{a}\right) \left|\vec{v}\left(\vec{a}, t\right)\right|^2 \, d\vec{a}.$$

Since R is orthogonal and $\vec{v} = R\vec{V}$, we have $|\vec{v}|^2 = |\vec{V}|^2$. Therefore, using (3.78), the kinetic energy of the body is given in terms of the body angular velocity by

$$T = \frac{1}{2} \int_{\mathcal{B}} \rho\left(\vec{a}\right) \left|\Omega(t) \times \vec{a}\right|^2 \, d\vec{a}. \tag{3.82}$$

From (3.80), this expression may be written as

$$T = \frac{1}{2}\vec{\Omega} \cdot \mathbf{I}\vec{\Omega}.$$

Thus, the body angular momentum $\vec{\Pi}$ is given by

$$\vec{\Pi} = \frac{\partial T}{\partial \vec{\Omega}}.$$

Note that this equation is dimensionally consistent, since $[T] = ML^2/T^2$. Expressed in terms of $\vec{\Pi}$, the kinetic energy is

$$T = \frac{1}{2}\vec{\Pi} \cdot \left(\mathbf{I}^{-1}\vec{\Pi}\right).$$

As we will show below, the kinetic energy T is conserved for solutions of (3.81).

12.4. The rigid body Poisson bracket

The equations (3.81) are a 3×3 system, so they cannot be canonical Hamiltonian equations, which are always even in number. We can, however, write them in Poisson form by use of a suitable noncanonical Poisson bracket.

We define a Poisson bracket of functions $F, G : \mathbb{R}^3 \to \mathbb{R}$ by

$$\{F, G\} = -\vec{\Pi} \cdot \left(\frac{\partial F}{\partial \vec{\Pi}} \times \frac{\partial G}{\partial \vec{\Pi}}\right), \tag{3.83}$$

This bracket is a skew-symmetric, bilinear derivation. It also satisfies the Jacobi identity (3.67), as may be checked by a direct computation. The minus sign is not required in order for (3.83) to define a bracket, but it is included to agree with the usual sign convention, which is related to a difference between right and left invariance in the Lie group $SO(3)$ underlying this problem.

For each $\vec{\Pi} \in \mathbb{R}^3$, we define a linear map $\mathbf{J}\left(\vec{\Pi}\right) : \mathbb{R}^3 \to \mathbb{R}^3$ by

$$\mathbf{J}\left(\vec{\Pi}\right)\vec{x} = \vec{\Pi} \times \vec{x}.$$

Then, using the cyclic symmetry of the scalar triple product, we may write the Poisson bracket as

$$\{F, G\} = \frac{\partial F}{\partial \vec{\Pi}} \cdot \mathbf{J}\left(\vec{\Pi}\right)\left[\frac{\partial G}{\partial \vec{\Pi}}\right],$$

Equation (3.81) is then

$$\dot{\vec{\Pi}} = \mathbf{J}\left(\vec{\Pi}\right)\left[\frac{\partial T}{\partial \vec{\Pi}}\right]$$

or, in Poisson bracket form,

$$\dot{\vec{\Pi}} = \left\{\vec{\Pi}, T\right\}, \qquad \dot{\Pi}_i = \{\Pi_i, T\}.$$

Next let us derive the conserved quantities for this equation Any function F such that $\{F, T\} = 0$ is conserved. In particular, since the Poisson bracket is skew-symmetric, the kinetic energy T itself is conserved.

Let $L : \mathbb{R}^3 \to \mathbb{R}$ denote the total angular momentum function

$$L\left(\vec{\Pi}\right) = \vec{\Pi} \cdot \vec{\Pi}.$$

Then, from (3.83),

$$\{F, L\} = -2\vec{\Pi} \cdot \left(\frac{\partial F}{\partial \vec{\Pi}} \times \vec{\Pi} \right) = -2 \frac{\partial F}{\partial \vec{\Pi}} \cdot \left(\vec{\Pi} \times \vec{\Pi} \right) = 0.$$

Thus, $\{F, L\} = 0$ for *any* function F. Such a function L is called a *Casimir* (or distinguished) function of the Poisson bracket; it is a conserved quantity for any Hamiltonian with that Poisson bracket. In particular, it follows that L is a conserved quantity for (3.81)

The conservation of L is also easy to derive directly from (3.81). Taking the inner product of the equation with π, we get

$$\frac{d}{dt} \left(\frac{1}{2} \vec{\pi} \cdot \vec{\pi} \right) = 0,$$

and, since R is orthogonal, $\vec{\pi} \cdot \vec{\pi} = \vec{\Pi} \cdot \vec{\Pi}$.

Thus, the trajectories of (3.81) lie on the intersection of the invariant spheres of constant angular momentum

$$\vec{\Pi} \cdot \vec{\Pi} = \text{constant}.$$

and the invariant ellipsoids of constant energy

$$\vec{\Pi} \cdot \left(\mathbf{I}^{-1} \vec{\Pi} \right) = \text{constant}.$$

To explain this picture in more detail, we write the rigid body equations in component form. Let $\{\vec{e}_1, \vec{e}_2, \vec{e}_3\}$ be an orthonormal basis of eigenvectors, or principal axes, of $\mathbf{I}$. There is such a basis because $\mathbf{I}$ is symmetric. We denote the corresponding eigenvalues, or principal moments of inertia, by $I_j > 0$, where

$$\mathbf{I} \vec{e}_j = I_j \vec{e}_j.$$

The eigenvalues are positive since $\mathbf{I}$ is positive definite. (It also follows from (3.80) that if $I_1 \leq I_2 \leq I_3$, say, then $I_3 \leq I_1 + I_2$.)

We expand

$$\vec{\Pi}(t) = \sum_{j=1}^{3} \Pi_j(t) \vec{e}_j$$

with respect to this principal axis basis. The component form of (3.81) is then

$$\begin{aligned}
\dot{\Pi}_1 &= \left(\frac{1}{I_3} - \frac{1}{I_2} \right) \Pi_2 \Pi_3, \\
\dot{\Pi}_2 &= \left(\frac{1}{I_1} - \frac{1}{I_3} \right) \Pi_3 \Pi_1, \\
\dot{\Pi}_3 &= \left(\frac{1}{I_2} - \frac{1}{I_1} \right) \Pi_1 \Pi_2.
\end{aligned}$$

Restricting this system to the invariant sphere

$$\Pi_1^2 + \Pi_2^2 + \Pi_3^2 = 1,$$

we see that there are three equilibrium points $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$, corresponding to steady rotations about each of the principle axes of the bodies. If $I_1 < I_2 < I_3$, then the middle equilibrium is an unstable saddle point, while the other two equilibria are stable centers.

The instability of the middle equilibrium can be observed by stretching an elastic band around a book and spinning it around each of its three axes.

This rigid-body Poisson bracket has a geometrical interpretation as a Poisson bracket on $\mathfrak{so}(3)^*$, the dual of the lie algebra of the three-dimensional rotation group $SO(3)$. Here, this dual space is identified with $\mathbb{R}^3$ through the cross-product and the Euclidean inner product.

There is an analogous Poisson bracket, called a Lie-Poisson bracket, on the dual of any Lie algebra. Like the rigid-body bracket, it depends linearly on the coordinates of the dual Lie algebra. Arnold observed that the equations of incompressible, inviscid fluid flows may be interpreted as Lie-Poisson equations associated with the infinite-dimensional group of volume-preserving diffeomorphisms on the fluid domain.

13. Hamiltonian PDEs

The Euler-Lagrange equation of a variational PDE can be transformed into a canonical Hamiltonian PDE in an analogous way to ODEs.

For example, consider the wave equation (3.57) with Lagrangian $\mathcal{L}(u, u_t)$ in (3.55). We define the momentum $p(\cdot, t)$, conjugate to the field variable $u(\cdot, t)$ by

$$p = \frac{\delta \mathcal{L}}{\delta u_t}$$

For (3.55), we get

$$p = \rho_0 u_t.$$

We then define the Hamiltonian functional $\mathcal{H}(u, p)$ by

$$\mathcal{H}(u, p) = \int p u_t \, dx - \mathcal{L}(u, u_t).$$

For (3.55), we get

$$\mathcal{H}(u, p) = \frac{1}{2} \int \left\{ \frac{p^2}{\rho_0} + k \left| \nabla u \right|^2 \right\} dx.$$

Hamilton's equations are

$$u_t = \frac{\delta \mathcal{H}}{\delta p}, \qquad p_t = -\frac{\delta \mathcal{H}}{\delta u}.$$

For (3.55), we find that

$$u_t = \frac{p}{\rho_0}, \qquad p_t = k \Delta u.$$

The elimination of p from this equation yields the wave equation (3.57).

The Poisson bracket of two functionals $\mathcal{F}(u, p)$, $\mathcal{G}(u, p)$ associated with these canonical variables is

$$\{\mathcal{F}, \mathcal{G}\} = \int \left(\frac{\delta \mathcal{F}}{\delta u} \frac{\delta \mathcal{G}}{\delta p} - \frac{\delta \mathcal{F}}{\delta p} \frac{\delta \mathcal{G}}{\delta u} \right) dx$$

Then, as before, for any functional $\mathcal{F}(u, p)$, evaluated on a solutions of Hamilton's equation, we have

$$\mathcal{F}_t = \{\mathcal{F}, \mathcal{H}\}.$$

13.1. Poisson brackets

One advantage of the Poisson bracket formulation is that it generalizes easily to PDE problems in which a suitable choice of canonical variables is not obvious.

Consider, for simplicity, an evolution equation that is first-order in time for a scalar-valued function $u(x,t)$. Suppose that $\mathbf{J}(u)$ is a skew-symmetric, linear operator on functions, which may depend upon u. In other words, this means that

$$\int f(x)\mathbf{J}(u)\,[g(x)]\;dx = -\int \mathbf{J}(u)\,[f(x)]\,g(x)\,dx.$$

for all admissible functions f, g, u. Here, we choose the integration range as appropriate; for example, we take the integral over all of space if the functions are defined on $\mathbb{R}$ or $\mathbb{R}^n$, or over a period cell if the functions are spatially periodic. We also assume that the boundary terms from any integration by parts can be neglected; for example, because the functions and their derivatives decay sufficiently rapidly at infinity, or by periodicity.

We then define a Poisson bracket of two spatial functionals $\mathcal{F}(u)$, $\mathcal{G}(u)$ of u by

$$\{\mathcal{F},\mathcal{G}\} = \int \frac{\delta\mathcal{F}}{\delta u}\cdot\mathbf{J}(u)\left[\frac{\delta\mathcal{G}}{\delta u}\right]\,dx$$

This bracket is a skew-symmetric derivation. If $\mathbf{J}$ is a constant operator that is independent of u, then the bracket satisfies the Jacobi identity (3.67), but, in general, the Jacobi identity places severe restrictions on how $\mathcal{J}(u)$ can depends on u (see [**40**]).

As an example, let us consider the Hamiltonian formulation of the KdV equation

$$u_t + uu_x + u_{xxx} = 0. \tag{3.84}$$

We define a constant skew-symmetric operator

$$\mathbf{J} = \partial_x$$

and a Hamiltonian functional

$$\mathcal{H}(u) = \int \left\{-\frac{1}{6}u^3 + \frac{1}{2}u_x^2\right\}$$

Then

$$\frac{\delta\mathcal{H}}{\delta u} = -\frac{1}{2}u^2 - u_{xx}$$

and hence the KdV equation (3.84) may be written as

$$u_t = \mathbf{J}\left[\frac{\delta\mathcal{H}}{\delta u}\right]$$

The associated Poisson bracket is

$$\{\mathcal{F},\mathcal{G}\} = \int \frac{\delta\mathcal{F}}{\delta u}\,\partial_x\left[\frac{\delta\mathcal{G}}{\delta u}\right]\,dx$$

The KdV equation is remarkable in that it has two different, but compatible, Hamiltonian formulations. This property is one way to understand the fact that the KdV equation is a completely integrable Hamiltonian PDE.

The second structure has the skew-symmetric operator

$$\mathbf{K}(u) = \frac{1}{3}\,(u\partial_x + \partial_x u) + \partial_x^3.$$

Note that the order of the operations here is important:

$$u\partial_x \cdot f = uf_x, \qquad \partial_x u \cdot f = (uf)_x = uf_x + u_x f.$$

Thus, the commutator of the multiplication operator u and the partial derivative operator ∂_x, given by $[u, \partial_x]\, f = -u_x f$, is the multiplication operator $-u_x$.

The Poisson bracket associated with $\mathbf{K}$ satisfies the Jacobi identity (this depends on a nontrivial cancelation). In fact, the Poisson bracket associated with $\alpha\mathbf{J} + \beta\mathbf{K}$ satisfies the Jacobi identity for any constants α, β, which is what it means for the Poisson structures to be compatible.

The KdV-Hamiltonian for $\mathbf{K}$ is

$$\mathcal{P}(u) = -\frac{1}{2}\int u^2\, dx,$$

with functional derivative

$$\frac{\delta \mathcal{P}}{\delta u} = -u.$$

The KdV equation may then be written as

$$u_t = \mathbf{K}\left[\frac{\delta \mathcal{P}}{\delta u}\right].$$

14. Path integrals

Feynman gave a remarkable formulation of quantum mechanics in terms of path integrals. The principle of stationary action for classical mechanics may be understood heuristically as arising from a stationary phase approximation of the Feynman path integral.

The method of stationary phase provides an asymptotic expansion of integrals with a rapidly oscillating integrand. Because of cancelation, the behavior of such integrals is dominated by contributions from neighborhoods of the stationary phase points where the oscillations are the slowest. Here, we explain the basic idea in the case of one-dimensional integrals. See Hormander [**27**] for a complete discussion.

14.1. Fresnel integrals

Consider the following Fresnel integral

$$I(\varepsilon) = \int_{-\infty}^{\infty} e^{ix^2/\varepsilon}\, dx. \tag{3.85}$$

This oscillatory integral is not defined as an absolutely convergent integral, since $e^{ix^2/\varepsilon}$ has absolute value one, but it can be defined as an improper Riemann integral

$$I(\varepsilon) = \lim_{R\to\infty} \int_{-R}^{R} e^{ix^2/\varepsilon}\, dx.$$

The convergence follows from an integration by parts:

$$\int_1^R e^{ix^2/\varepsilon}\, dx = \Big[\frac{\varepsilon}{2ix} e^{ix^2/\varepsilon}\Big]_1^R + \int_1^R \frac{\varepsilon}{2ix^2} e^{ix^2/\varepsilon}\, dx.$$

The integrand in (3.85) oscillates rapidly away from the stationary phase point $x = 0$, and these parts contribute terms that are smaller than any power of ε as $\varepsilon \to 0$, as we show below. The first oscillation near $x = 0$, where cancelation does not occur, has width of the order $\varepsilon^{1/2}$, and as a result $I(\varepsilon) = O(\varepsilon^{1/2})$ as $\varepsilon \to 0$.

Using contour integration, and changing variables $x \mapsto e^{i\pi/4}s$ if $\varepsilon > 0$ or $x \mapsto e^{-i\pi/4}s$ if $\varepsilon < 0$, one can show that

$$\int_{-\infty}^{\infty} e^{ix^2/\varepsilon}\,dx = \begin{cases} e^{i\pi/4}\sqrt{2\pi|\varepsilon|} & \text{if } \varepsilon > 0, \\ e^{-i\pi/4}\sqrt{2\pi|\varepsilon|} & \text{if } \varepsilon < 0. \end{cases}$$

14.2. Stationary phase

Next, we consider the integral

$$I(\varepsilon) = \int_{-\infty}^{\infty} f(x)e^{i\varphi(x)/\varepsilon}\,dx, \tag{3.86}$$

where $f : \mathbb{R} \to \mathbb{C}$ and $\varphi : \mathbb{R} \to \mathbb{R}$ are smooth functions. A point $x = c$ is a stationary phase point if $\varphi'(c) = 0$. We call the stationary phase point nondegenerate if $\varphi''(c) \neq 0$.

Suppose that I has a single stationary phase point at $x = c$, and it is nondegenerate. If there are several such points, we simply add together the contributions from each one. Then, using the idea that only the part of the integrand near the stationary phase point $x = c$ contributes significantly, we Taylor expand the function f and the phase φ to approximate $I(\varepsilon)$ as follows:

$$\begin{aligned} I(\varepsilon) &\sim \int f(c)\exp\frac{i}{\varepsilon}\left[\varphi(c) + \frac{1}{2}\varphi''(c)(x-c)^2\right]\,dx \\ &\sim f(c)e^{i\varphi(c)/\varepsilon}\int \exp\left[\frac{i\varphi''(c)}{2\varepsilon}s^2\right]\,ds \\ &\sim \sqrt{\frac{2\pi\varepsilon}{|\varphi''(c)|}}\,f(c)e^{i\varphi(c)/\varepsilon + i\sigma\pi/4}, \end{aligned}$$

where

$$\sigma = \operatorname{sign}\varphi''(c).$$

14.3. The Feynman path integral

Consider a single, non-relativistic quantum mechanical particle of mass m in a potential $V(\vec{x})$. Suppose that the particle is located at $\vec{x}_0$ at time t_0, and we observe the location of the particle at time t_1. We would like to calculate the probability of finding the particle in some specific region $\Omega \subset \mathbb{R}^n$.

According to Feynman's formulation of quantum mechanics [**21**], every event has an associated complex number, Ψ, called its amplitude. If an event can occur in a number of different independent ways, the amplitude of the event is obtained by adding together the amplitudes of the different subevents. Finally, the probability of observing an event when some measurement is made is the modulus of the amplitude squared $|\Psi|^2$.

The fact that amplitudes add, not probabilities, leads to the interference effects characteristic of quantum mechanics. For example, consider an event (like the observation of an electron in the 'double slit' experiment) which can occur in two different ways with equal probability. If the two amplitudes have opposite phase, then the probability of the event is zero, while if they have the same phase, then the probability of the event is four times the probability of the separate subevents.

To apply this formulation to the motion of a quantum mechanical particle, we take as the basic subevents the possible paths $\vec{x}(t)$ of the particle from $\vec{x}_0$ at time

t_0 to $\vec{x}_1$ at time t_1. The amplitude of a path $\vec{x}$ is proportional to $e^{i\mathcal{S}(\vec{x})/\hbar}$ where $\mathcal{S}(\vec{x})$ is the action of the path

$$\mathcal{S}(\vec{x}) = \int_{t_0}^{t_1} \left\{ \frac{1}{2} m \left| \dot{\vec{x}} \right|^2 - V(\vec{x}) \right\} dt,$$

and $\hbar$ is Planck's constant. Like the action, Planck's constant has the dimension of energy $\cdot$ time, or momentum $\cdot$ length; its approximate value is $\hbar = 1.054 \times 10^{-34}$ Js.

Thus the action, which is a somewhat mysterious quantity in classical mechanics, corresponds to a phase, measured in units of $\hbar$, in quantum mechanics.

The amplitude $\psi(\vec{x}_1, t_1; \vec{x}_0, t_0)$ of the particle moving from $\vec{x}_0$ at time t_0 to $\vec{x}_1$ at time t_1 is then obtained formally by summing the amplitudes of each path over 'all' possible paths

$$\mathcal{P}(\vec{x}_1, t_1; \vec{x}_0, t_0) = \{\vec{x} \mid \vec{x}(t) : [t_0, t_1] \to \mathbb{R}^n \text{ is continuous, } \vec{x}(t_0) = \vec{x}_0, \vec{x}(t_0) = \vec{x}_0\}.$$

This gives

$$\psi(\vec{x}_1, t_1; \vec{x}_0, t_0) = \int_{\mathcal{P}(\vec{x}_1, t_1; \vec{x}_0, t_0)} e^{i\mathcal{S}(\vec{x})/\hbar} \, D\vec{x}, \tag{3.87}$$

where $D\vec{x}$ is supposed to be a measure on the path space that weights all paths equally, normalized so that $|\psi|^2$ is a probability density.

This argument has great intuitive appeal, but there are severe difficulties in making sense of the result. First, there is no translation-invariant 'flat' measure $D\vec{x}$ on an infinite-dimensional path space, analogous to Lebesgue measure on $\mathbb{R}^n$, that weights all paths equally. Second, for paths $\vec{x}(t)$ that are continuous but not differentiable, which include the paths one expects to need, the action $\mathcal{S}(\vec{x})$ is undefined, or, at best, infinite. Thus, in qualitative terms, the Feynman path integral in the expression for ψ in fact looks something like this:

$$\text{“}\textstyle\int_{\mathcal{P}} e^{i\mathcal{S}(\vec{x})/\hbar} \, D\vec{x} = \int e^{i\infty/\hbar} \, D?\text{”}.$$

Nevertheless, there are ways to make sense of (3.87) as providing the solution $\psi(\vec{x}, t; \vec{x}_0, t_0)$ of the Schrödinger equation

$$\begin{aligned} i\hbar\psi_t &= -\frac{\hbar^2}{2m}\Delta\psi + V(\vec{x})\,\psi, \\ \psi(\vec{x}, t_0; \vec{x}_0, t_0) &= \delta(\vec{x} - \vec{x}_0). \end{aligned}$$

For example, the Trotter product formula gives an expression for ψ as a limit of finite dimensional integrals over $\mathbb{R}^N$ as $N \to \infty$, which may be taken as a definition of the path integral in (3.87) (see (3.92)–(3.94) below).

After seeing Feyman's work, Kac (1949) observed that an analogous formula for solutions of the heat equation with a lower-order potential term (the 'imaginary time' version of the Schrödinger equation),

$$u_t = \frac{1}{2}\Delta u - V(x)u, \tag{3.88}$$

can be given rigorous sense as a path integral with respect to Wiener measure, which describes the probability distribution of particle paths in Brownian motion.

Explicitly, for sufficiently smooth potential functions $V(x)$, the Green's function of (3.88), with initial data $u(x, t_0) = \delta(x - x_0)$, is given by the Feynman-Kac formula

$$u(x, t; x_0, t_0) = \int_{\mathcal{P}(x, t; x_0, t_0)} e^{\int_{t_0}^{t} V(x(s))\, ds} \, dW(x).$$

Here, $\mathcal{P}(x,t;x_0,t_0)$ denotes the space of all continuous paths $x(s)$, with $t_0 \leq s \leq t$ from x_0 at t_0 to x at t. The integral is taken over $\mathcal{P}$ with respect to Wiener measure. Formally, we have

$$dW(x) = e^{-\int_{t_0}^{t} \frac{1}{2}\left|\dot{\vec{x}}\right|^2 ds} Dx. \tag{3.89}$$

However, neither the 'flat' measure Dx, nor the exponential factor on the right-hand side of this equation are well-defined. In fact, the Wiener measure is supported on continuous paths that are almost surely nowhere differentiable (see Section 3.4).

14.4. The Trotter product formula

To explain the idea behind the Trotter product formula, we write the Schrödinger equation as

$$i\hbar\psi_t = H\psi, \tag{3.90}$$

where the Hamiltonian operator H is given by

$$H = T + V$$

and the kinetic and potential energy operators T and V, respectively, are given by

$$T = -\frac{\hbar^2}{2m}\Delta, \qquad V = V(\vec{x}).$$

Here, V is understood as a multiplication operator $V : \psi \mapsto V(\vec{x})\,\psi$.

We write the solution of (3.90) as

$$\psi(t) = e^{-it\hbar^{-1}H}\psi_0$$

where $\psi(t) = \psi(\cdot,t) \in L^2(\mathbb{R}^n)$ denotes the solution at time t, $\psi_0 = \psi(0)$ is the initial data, and

$$e^{-it\hbar^{-1}H} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$$

is the one-parameter group of solution operators (or flow).

Assuming that $V(\vec{x})$ is not constant, the operators T, V do not commute:

$$[T,V] = TV - VT = -\frac{\hbar^2}{2m}\left(2\nabla V\cdot\nabla + \Delta V\right).$$

(Here, ΔV denotes the operation of multiplication by the function ΔV.) Thus, the flows $e^{-it\hbar^{-1}T}$, $e^{-it\hbar^{-1}V}$ do not commute, and $e^{-it\hbar^{-1}H} \neq e^{-it\hbar^{-1}V}e^{-it\hbar^{-1}T}$.

For small times Δt, however, we have

$$\begin{aligned}
e^{-i\Delta t\hbar^{-1}H} &= I - \frac{i\Delta t}{\hbar}H - \frac{\Delta t^2}{2\hbar^2}H^2 + O\left(\Delta t^3\right) \\
&= I - \frac{i\Delta t}{\hbar}(T+V) - \frac{\Delta t^2}{2\hbar^2}\left(T^2 + TV + VT + V^2\right) + O\left(\Delta t^3\right), \\
e^{-i\Delta t\hbar^{-1}T} &= I - \frac{i\Delta t}{\hbar}T - \frac{\Delta t^2}{2\hbar^2}T^2 + O\left(\Delta t^3\right) \\
e^{-i\Delta t\hbar^{-1}V} &= I - \frac{i\Delta t}{\hbar}V - \frac{\Delta t^2}{2\hbar^2}V^2 + O\left(\Delta t^3\right).
\end{aligned}$$

Thus,

$$e^{-i\Delta t\hbar^{-1}H} = e^{-i\Delta t\hbar^{-1}V}e^{-i\Delta t\hbar^{-1}T} - \frac{\Delta t^2}{2\hbar^2}[T,V] + O\left(\Delta t^3\right),$$

and we can obtain a first-order accurate approximation for the flow associated with H by composing the flows associated with V and T.

The numerical implementation of this idea is the fractional step method. We solve the evolution equation

$$u_t = (A + B)u$$

by alternately solving the equations

$$u_t = Au, \qquad u_t = Bu$$

over small time-steps Δt. In this context, the second-order accurate approximation in Δt

$$e^{\Delta t(A+B)} = e^{\frac{1}{2}\Delta t A} e^{\Delta t B} e^{\frac{1}{2}\Delta t A}$$

is called 'Strang splitting.'

To obtain the solution of (3.90) at time t, we take N time-steps of length $\Delta t = t/N$, and let $N \to \infty$, which gives the Trotter product formula

$$\psi(t) = \lim_{N\to\infty} \left[e^{-it(\hbar N)^{-1} V} e^{-it(\hbar N)^{-1} T} \right]^N \psi_0. \tag{3.91}$$

Under suitable assumptions on V, the right-hand side converges strongly to $\psi(t)$ with respect to the $L^2(\mathbb{R}^n)$-norm.

The flows associated with V, T are easy to find explicitly. The solution of

$$i\hbar\psi_t = V\psi$$

is given by the multiplication operator

$$\psi_0 \mapsto e^{-it\hbar^{-1}V}\psi_0.$$

The solution of

$$i\hbar\psi_t = T\psi$$

may be found by taking the spatial Fourier transform and using the convolution theorem, which gives

$$\begin{aligned} \psi(\vec{x}, t) &= \int e^{\left\{-it|\vec{p}|^2/(2\hbar m) + i\vec{p}\cdot\vec{x}/\hbar\right\}} \hat{\psi}_0(\vec{p})\, d\vec{p} \\ &= \left(\frac{m}{2\pi i\hbar t}\right)^{n/2} \int e^{im|\vec{x}-\vec{y}|^2/(2\hbar t)} \psi_0(\vec{y})\, d\vec{y}. \end{aligned}$$

Using these results in the Trotter product formula (3.91), writing the spatial integration variable at time $t_k = kt/N$ as $\vec{x}_k$, with $\vec{x}_N = \vec{x}$, and assuming that $\psi(\vec{x}, 0) = \delta(\vec{x} - \vec{x}_0)$, we get, after some algebra, that

$$\psi(\vec{x}, t) = \lim_{N\to\infty} C_{N,t} \int e^{iS_{N,t}(\vec{x}_0, \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_{N-1}, \vec{x}_N)/\hbar}\, d\vec{x}_1 d\vec{x}_2, \ldots d\vec{x}_{N-1} \tag{3.92}$$

where the normalization factor $C_{N,t}$ is given by

$$C_{N,t} = \left(\frac{mN}{2\pi i\hbar t}\right)^{n(N-1)/2} \tag{3.93}$$

and the exponent $S_{N,t}$ is a discretization of the classical action functional

$$S_{N,t}(\vec{x}_0, \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_{N-1}, \vec{x}_N) = \sum_{k=1}^{N-1} \frac{t}{N} \left[\frac{m}{2} \left| \frac{\vec{x}_{k+1} - \vec{x}_k}{t/N} \right|^2 - V(\vec{x}_k) \right]. \tag{3.94}$$

Equations (3.92)–(3.94) provide one way to interpret the path integral formula (3.87).

14.5. Semiclassical limit

One of the most appealing features of the Feynman path integral formulation is that it shows clearly the connection between classical and quantum mechanics. The phase of the quantum mechanical amplitude is the classical action, and, by analogy with the method of stationary phase for finite-dimensional integrals, we expect that for semi-classical processes whose actions are much greater than $\hbar$, the amplitude concentrates on paths of stationary phase. Again, however, it is difficult to make clear analytical sense of this argument while maintaining its simple intuitive appeal.

LECTURE 4

Sturm-Liouville Eigenvalue Problems

Possibly one of the most useful facts in mathematics is that a symmetric matric has real eigenvalues and a set of eigenvectors that form an orthonormal basis. This property of symmetric matrices has a wide-ranging generalization to the spectral properties of self-adjoint operators in a Hilbert space, of which the Sturm-Liouville ordinary differential operators are fundamental examples.

Sturm-Liouville equations arise throughout applied mathematics. For example, they describe the vibrational modes of various systems, such as the vibrations of a string or the energy eigenfunctions of a quantum mechanical oscillator, in which case the eigenvalues correspond to the resonant frequencies of vibration or energy levels. It was, in part, the idea that the discrete energy levels observed in atomic systems could be obtained as the eigenvalues of a differential operator which led Schrödinger to propose his wave equation.

Sturm-Liouville problems arise directly as eigenvalue problems in one space dimension. They also commonly arise from linear PDEs in several space dimensions when the equations are separable in some coordinate system, such as cylindrical or spherical coordinates.

The general form of the Sturm-Liouville equation is an ODE for $u(x)$ of the form

$$-(pu')' + qu = \lambda ru. \tag{4.1}$$

Here, $p(x)$, $q(x)$ are coefficient functions, $r(x)$ is a weighting function (equal to one in the simplest case) and λ is an eigenvalue, or spectral, parameter. The ODE is supplemented by appropriate self-adjoint boundary conditions.

The simplest example of a Sturm-Liouville operator is the constant-coefficient second-derivative operator, whose eigenfunctions are trigonometric functions. Many other important special functions, such as Airy functions and Bessel functions, are associated with variable-coefficient Sturm-Liouville operators.

Just as we may expand a vector with respect to the eigenvectors of a symmetric matrix, we may expand functions in terms of the eigenfunctions of a regular Sturm-Liouville operator; the expansion of periodic functions in Fourier series is an example.

One feature that occurs for Sturm-Liouville operators, which does not occur for matrices, is the possibility of an absolutely continuous (or, for short, continuous) spectrum. Instead of eigenfunction expansions, we then then get integral transforms, of which the Fourier transform is an example.

Other, more complicated spectral phenomena can also occur. For example, eigenvalues embedded in a continuous spectrum, singular continuous spectrum, and pure point spectrum consisting of eigenvalues that are dense in an interval (see Section 4.6 on the Anderson localization of waves in random media for an example).

1. Vibrating strings

Consider the vibrations of a string such as a violin string. We label material points on the string by a Lagrangian coordinate $a \in \mathbb{R}$; for example, we can define a as the distance of the point from some fixed point in a given reference configuration of the string. We denote the position of the material point a on the string at time t by $\vec{r}(a,t)$.

Let $\rho_0(a)$ denote the mass-density of the string in the reference configuration, meaning that the mass of the part of the string with $c \le a \le d$ is given by

$$\int_c^d \rho_0(a)\, da.$$

We assume that the mass of the string is conserved as it vibrates, in which case the density $\rho_0(a)$ in the reference configuration is independent of time.

We suppose that the only force exerted by one part of the string on another is a tension force tangent to the string. This assumption distinguishes a string from an elastic rod that resists bending. We also suppose, for simplicity, that no external forces act on the string.

The contact force $\vec{F}$ exerted by the part of the string with $b > a$ on the part with $b < a$ is then given by

$$\vec{F}(a,t) = T(a,t)\,\vec{t}(a,t)$$

where $T(a,t)$ is the tension in the string and

$$\vec{t} = \frac{\vec{r}_a}{|\vec{r}_a|}$$

is the unit tangent vector. We assume that $\vec{r}_a$ never vanishes. The part of the string with $b < a$ exerts an equal and opposite force $-\vec{F}$ on the part of the string with $b > a$.

Newton's second law, applied to the part of the string with $c \le a \le d$, gives

$$\frac{d}{dt}\int_c^d \rho_0(a)\vec{r}_t(a,t)\, da = \vec{F}(d,t) - \vec{F}(c,t).$$

For smooth solutions, we may rewrite this equation as

$$\int_c^d \left\{\rho_0(a)\vec{r}_{tt}(a,t) - \vec{F}_a(a,t)\right\}\, da = 0.$$

Since this holds for arbitrary intervals $[c,d]$, and since we assume that all functions are smooth, we conclude that

$$\rho_0(a)\vec{r}_{tt}(a,t) = \vec{F}_a(a,t).$$

This equation expresses conservation of momentum for motions of the string.

To close the equation, we require a constitutive relation that relates the tension in the string to the stretching of the string. The local extension of the string from its reference configuration is given by

$$e(a,t) = |\vec{r}_a(a,t)|.$$

We assume that

$$T(a,t) = f\left(e(a,t), a\right)$$

where $f(e,a)$ is a given function of the extension e and the material coordinate a.

It follows that the position-vector $\vec{r}(a,t)$ of the string satisfies the nonlinear wave equation

$$\rho_0(a)\vec{r}_{tt}(a,t) = \partial_a \left\{ f\left(|\vec{r}_a(a,t)|, a\right) \frac{\vec{r}_a(a,t)}{|\vec{r}_a(a,t)|} \right\}. \tag{4.2}$$

1.1. Equilibrium solutions

A function $\vec{r} = \vec{r}_0(a)$ is an exact, time-independent solution of (4.2) if

$$\frac{d}{da} \left\{ f\left(|\vec{r}_{0a}|, a\right) \frac{\vec{r}_{0a}}{|\vec{r}_{0a}|} \right\} = 0. \tag{4.3}$$

We consider a solution such that the tangent vector of the string is in a constant direction, say the $\vec{i}$-direction.

We may then use as a material coordinate the distance a along the string in the equilibrium configuration, in which case

$$\vec{r}_0(a) = a\vec{i}. \tag{4.4}$$

Using (4.4) and the corresponding extension $e = 1$, in (4.3), we find that the tension

$$f\left(1, a\right) = T_0$$

is constant in equilibrium, as required by the balance of longitudinal forces.

1.2. Linearized equation

For small vibrations of the string about an equilibrium state, we may linearize the equations of motion. We look for solutions of the form

$$\vec{r}(a,t) = a\vec{i} + \vec{r}'(a,t), \tag{4.5}$$

where $\vec{r}'$ is a small perturbation of the equilibrium solution (4.4). We decompose $\vec{r}'$ into longitudinal and transverse components

$$\vec{r}'(a,t) = x'(a,t)\vec{i} + \vec{r}^{\perp}(a,t),$$

where $\vec{i} \cdot \vec{r}^{\perp} = 0$.

We use (4.5) in (4.2), with $e = 1 + x'_a + \ldots$, and Taylor expand the resulting equation with respect to $\vec{r}'$. This gives

$$\rho_0 \vec{r}'_{tt} = \partial_a \left\{ \left(T_0 + kx'_a\right)\left(1 - x'_a\right) \left[\left(1 + x'_a\right)\vec{i} + \vec{r}^{\perp}_a \right] \right\} + \ldots$$

where $k(a) = f_e\left(1, a\right)$. Linearizing the equation, and separating it into longitudinal and transverse components, we find that

$$\rho_0 x'_{tt} = \left(kx'_a\right)_a, \qquad \rho_0 \vec{r}^{\perp}_{tt} = T_0 \vec{r}^{\perp}_{aa}. \tag{4.6}$$

Thus we obtain decoupled equations for the longitudinal and transverse motions of the string.

The longitudinal displacement satisfies a one-dimensional wave equation of the form (3.56). The density is given by the density in the reference configuration, and the stiffness by the derivative of the tension with respect to the extension; the stiffness is positive, and the equation is a wave equation provided that the tension in the string increases when it is stretched. In general, both coefficients are functions of a, but for a uniform string they are constants.

Unlike the longitudinal mode, the stiffness constant T_0 for the transverse mode is necessarily constant. If $T_0 > 0$, meaning that the string is stretched, the transverse displacement satisfies a wave equation, but if $T_0 < 0$ it satisfies an elliptic

equation. As we explain in the following section, the initial value problem for such PDEs is subject to an extreme instability. This is consistent with our experience that one needs to stretch a string to pluck it.

1.3. Hadamard instability

As a short aside, we consider the instability of the initial value problem for an elliptic PDE, such as the one that arises above for transverse vibrations of a compressed elastic string. This type of instability arises in other physical problems, such as the Kelvin-Helmholtz instability of a vortex sheet in fluid mechanics.

The simplest case of the transverse equation in (4.6) with constant coefficients, normalized to $\rho_0 = 1$, $T_0 = -1$, and planar motions $x = a$ and $y = u(x,t)$, is the Laplace equation

$$u_{tt} = -u_{xx}. \tag{4.7}$$

Equation (4.7) has solutions

$$u(x,t) = Ae^{inx+|n|t} \tag{4.8}$$

for arbitrary $n \in \mathbb{R}$ and $A \in \mathbb{C}$. Since the equation is linear with real-coefficients, we may obtain real-valued solutions by taking the real or imaginary parts of any complex-valued solution, and we consider complex-valued solutions for convenience.

The solution in (4.8) has modulus $|u(x,t)| = |A|e^{|k|t}$. Thus, these solutions grow exponentially in time with arbitrarily large rates. (The solutions proportional to $e^{inx-|n|t}$ grow arbitrarily fast backward in time.)

This behavior is a consequence of the invariance of (4.8) under the rescalings $x \mapsto \lambda$, $t \mapsto \lambda t$. This scale-invariance implies that if there is one solution with bounded initial data and a nonzero growth rate, then we can obtain solutions with arbitrarily fast growth rates by rescaling the initial data.

As a result, solutions do not depend continuously on the initial data in any norm that involves only finitely many derivatives, and the resulting initial value problem for (4.7) is ill-posed with respect to such norms. For example, if

$$\|u\|_K(t) = \max_{0\leq k\leq K} \sup_{x\in\mathbb{R}} \left|\partial_x^k u(x,t)\right|$$

and

$$u_n(x,t) = e^{-|n|^{1/2}} \left\{e^{inx-nt} + e^{inx+nt}\right\},$$

then for every $K \in \mathbb{N}$ we have

$$\|u_n\|_K(0) \to 0 \qquad \text{as } n \to \infty,$$

but

$$|u_n(x,t)| \to \infty \qquad \text{as } n \to \infty \text{ if } t \neq 0.$$

This failure of continuous dependence leads to a loss of existence of solutions. For example, the Fourier series

$$f(x) = \sum_{k=-\infty}^{\infty} e^{-|k|^{1/2}} e^{ikx}$$

converges to a C^∞-function, but there is no solution of (4.7) with initial data

$$u(x,0) = f(x), \qquad u_t(x,0) = 0$$

in any time interval about 0, however short.

It is possible to obtain solutions of (4.7) for sufficiently 'good' initial data, such as analytic functions (which, from the Cauchy-Kovalewsky theorem, are given by locally convergent power series). Such assumptions, however, are almost always too restrictive in applications. The occurrence of Hadamard instability typically signals a failure of the model, and means that additional stabilizing effects must be included at sufficiently short length-scales.

2. The one-dimensional wave equation

Consider a uniform string with constant density ρ_0 and and constant stiffness k_0. Then, from (4.17), longitudinal vibrations of the string satisfy the one-dimensional wave equation

$$u_{tt} = c_0^2 u_{xx} \tag{4.9}$$

Planar transverse vibrations of a stretched string satisfy the same equation with $c_0^2 = T_0/\rho_0$.

2.1. The d'Alembert solution

The general solution of (4.9) is given by the d'Alembert solution

$$u(x,t) = F(x - c_0 t) + G(x + c_0 t) \tag{4.10}$$

where F, G are arbitrary functions. This solution represents a superposition of a right-moving traveling wave with profile F and a left-moving traveling wave with profile G.

It follows that the solution of the Cauchy problem for (4.9) with initial data

$$u(x,0) = f(x), \qquad u_t(x,0) = g(x) \tag{4.11}$$

is given by (4.10) with

$$F(x) = \frac{1}{2} f(x) - \frac{1}{2c_0} \int_{x_0}^{x} g(\xi)\, d\xi, \qquad G(x) = \frac{1}{2} f(x) + \frac{1}{2c_0} \int_{x_0}^{x} g(\xi)\, d\xi.$$

Here, x_0 is an arbitrary constant; changing x_0 does not change the solution, it simply transforms $F(x) \mapsto F(x) + c$, $G(x) \mapsto G(x) - c$ for some constant c.

2.2. Normal modes

Next, consider a boundary value problem (BVP) for (4.9) in $0 \le x \le L$ with boundary conditions

$$u(0,t) = 0, \qquad u(L,t) = 0. \tag{4.12}$$

This BVP describes the vibration of a uniform string of length L that is pinned at its endpoints.

We look for separable solutions of the form

$$u(x,t) = \varphi(x) e^{-i\omega t} \tag{4.13}$$

where ω is a constant frequency and $\varphi(x)$ is a function of the spatial variable only. The real and imaginary parts of a complex-valued solution of a linear equation with real coefficients are also solutions, so we may recover the real-valued solutions from these complex-valued solutions.

The function $u(x,t)$ in (4.13) satisfies (4.9), (4.12) if $\varphi(x)$ satisfies

$$\varphi'' + k^2 \varphi = 0, \qquad \varphi(0) = 0, \quad \varphi(L) = 0$$

where the prime denotes a derivative with respect to x, and

$$k^2 = \frac{\omega^2}{c_0^2}.$$

The spectral problem

$$-\varphi'' = \lambda\varphi, \qquad \varphi(0) = 0, \quad \varphi(L) = 0$$

has a point spectrum consisting entirely of eigenvalues

$$\lambda_n = \frac{\pi^2 n^2}{L^2} \qquad \text{for } n = 1, 2, 3, \ldots. \tag{4.14}$$

Up to an arbitrary constant factor, the corresponding eigenfunctions $\varphi_n \in L^2[0, L]$ are given by

$$\varphi_n(x) = \sin\left(\frac{n\pi x}{L}\right). \tag{4.15}$$

These eigenfunctions are orthogonal with respect to the $L^2[0, L]$-inner product

$$\langle f, g\rangle = \int_0^L f(x)g(x)\,dx,$$

where $f, g : [0, L] \to \mathbb{R}$ are square-integrable functions. Explicitly, for any $m, n \in \mathbb{N}$

$$\int_0^L \sin\left(\frac{m\pi x}{L}\right)\sin\left(\frac{n\pi x}{L}\right)\,dx = \begin{cases} L/2 & \text{if } n = m, \\ 0 & \text{if } n \neq m. \end{cases}$$

An arbitrary function $f \in L^2[0, L]$ may be expanded with respect to these eigenfunctions in a Fourier sine series as

$$f(x) = \sum_{n=1}^{\infty} a_n \sin n\pi x,$$

where, by orthogonality,

$$a_n = \frac{2}{L}\int_0^L f(x)\sin n\pi x\,dx.$$

The series converges to f in the L^2-norm, meaning that

$$\left\| f(x) - \sum_{n=1}^{N} a_n \sin n\pi x \right\| = \left(\int_0^L \left| f(x) - \sum_{n=1}^{N} a_n \sin n\pi x \right|^2 dx \right)^{1/2} \to 0 \qquad \text{as } N \to \infty.$$

We say that the eigenfunctions are a basis of $L^2[0, L]$, and form a complete set.

The solutions for k corresponding to (4.14) are

$$k_n = \frac{n\pi}{L} \qquad \text{for } n = 1, 2, 3, \ldots,$$

The separable solutions of (4.9), (4.12) associated with (4.15) are therefore

$$\sin(k_n x)\, e^{-ic_0 k_n t}, \qquad \sin(k_n x)\, e^{ic_0 k_n t}.$$

The real part of these solutions is

$$u(x, t) = \sin(k_n x)\cos\left(c_0 k_n t\right). \tag{4.16}$$

This is a standing-wave solution with profile proportional to $\sin\left(k_n x\right)$ that oscillates periodically in time with frequency $\omega_n = c_0 n$. The n^{th} mode has $n/2$ periods of the sine that fit between the two pinned endpoints at $x = 0$, $x = L$.

The frequencies of these solutions are $\omega_n = c_0 k_n$. When expressed in terms of the properties of the string, the lowest, or fundamental, frequency for transverse modes is

$$\omega_1 = \sqrt{\frac{T_0}{\rho_0 L^2}}.$$

Thus, for example, increasing the tension in a string increases the fundamental frequency of its transverse vibrations, and heavier, longer strings have a lower fundamental frequency than lighter, shorter ones.

The solution (4.16) may be written as

$$u(x,t) = \frac{1}{2}\sin\left[n\pi\left(x - c_0 t\right)\right] + \frac{1}{2}\sin\left[n\pi\left(x + c_0 t\right)\right],$$

which shows that the standing wave arise from the interference between two traveling waves propagating in opposite directions.

Since the PDE and the boundary conditions (4.9), (4.12) are linear, we can superpose the separated, time-periodic solutions to get the general real-valued solution

$$u(x,t) = \sum_{n=1}^{\infty} \sin(n\pi x)\left\{A_n e^{ic_0 nt} + \bar{A}_n e^{-ic_0 nt}\right\}$$

where $A_n \in \mathbb{C}$ is an arbitrary constant for each $n \in \mathbb{N}$.

We may write this solution in real form as

$$u(x,t) = \sum_{n=1}^{\infty} \sin(n\pi x)\left\{a_n \cos\left(c_0 nt\right) + b_n \sin\left(c_0 nt\right)\right\}$$

where $A_n = (a_n + ib_n)/2$. Imposing the initial condition (4.11), we find that

$$a_n = 2\int_0^1 f(x)\sin(n\pi x)\,dx, \qquad b_n = \frac{2}{nc_0}\int_0^1 g(x)\sin(n\pi x)\,dx.$$

This solution can again be written as a superposition of right-moving and right-moving traveling waves.

Similar solutions can be obtained for Neumann boundary conditions

$$\varphi'(0) = 0, \qquad \varphi'(L) = 0$$

leading to Fourier cosine series, and periodic boundary conditions

$$\varphi(0) = \varphi(L), \qquad \varphi'(0) = \varphi'(L)$$

leading to Fourier series.

2.3. The Fourier transform

On the real line, the spectral problem

$$-\varphi'' = \lambda\varphi$$

has a continuous spectrum $0 \le \lambda < \infty$, with bounded solutions that do not lie in $L^2(\mathbb{R})$; they are linear combinations of $e^{\pm i\sqrt{\lambda}x}$. Since $k^2 = \lambda$ and $\omega^2 = c_0^2 k^2$, we get a continuous set of solutions of the wave equation, proportional to

$$u(x,t) = e^{ikx - ic_0 kt}, \qquad u(x,t) = e^{ikx + ic_0 kt}$$

where $k \in \mathbb{R}$. The general superposition of these solutions is

$$u(x,t) = \int \left\{\hat{F}(k)e^{ik(x-c_0 t)} + \hat{G}(k)e^{ik(x+c_0 t)}\right\} dk$$

where $\hat{F}, \hat{G} : \mathbb{R} \to \mathbb{C}$ are arbitrary functions. The solution $u(x,t)$ is real-valued if $\hat{F}(-k) = \hat{F}^*(k)$ and $\hat{G}(-k) = \hat{G}^*(k)$ This solution is the Fourier transform of the d'Alembert solution (4.10).

This solution exhibits an important feature of the wave equation, namely that it is nondispersive. Fourier modes with different wavenumbers k propagate with the same velocity c_0 (or $-c_0$). As a result, The Fourier modes stay together as the solution evolves in time, and the solution is a superposition of traveling wave solutions with arbitrary wave-profiles that propagate at velocities c_0 (or $-c_0$) without changing their shape.

This behavior contrast with the behavior of linear dispersive waves, where the velocity of the Fourier modes depends on their wavenumbers. In this case, the Fourier modes making up the initial data separate, or disperse, as the solution evolves, leading to an extended oscillatory wavetrain.

2.4. Nonuniform strings

Let us return to the one-dimensional wave equation

$$\rho_0 u_{tt} = (k u_x)_x \tag{4.17}$$

for the longitudinal displacement $u(x,t)$ of a string of length L with mass-density $\rho_0(x)$ and stiffness $k(x)$, both of which we assume are smooth, strictly positive functions on $0 \le x \le L$. Suppose, for definiteness, that $u(x,t)$ satisfies the homogeneous Dirichlet condition (4.12) at the endpoints of the string.

Looking for separable time-periodic solutions of (4.17) and (4.12) of the form (4.13), we get the BVP

$$\begin{aligned} -(k\varphi')' &= \lambda \rho_0 \varphi, \\ \varphi(0) = 0, &\qquad \varphi(L) = 0, \end{aligned}$$

where $\lambda = \omega^2$. This equation has the form of a Sturm-Liouville eigenvalue problem (4.1) with $p = k$, $q = 0$, and $r = \rho_0$. Values of ω for which this BVP has nonzero solutions correspond to resonant frequencies of oscillation of the string. Unlike the uniform case, we cannot solve this eigenvalue problem explicitly for general coefficient functions ρ_0, k.

The qualitative behavior is, however, the same. There is an infinite sequence of simple positive eigenvalues $\lambda_1 < \lambda_2 < \lambda_3 < \dots$. The corresponding eigenfunctions $\{\varphi_1, \varphi_2, \varphi_3, \dots\}$ are orthogonal with respect to the weighted inner-product

$$\langle f, g \rangle = \int_0^L \rho_0(x) f(x) g(x)\, dx,$$

and form a complete set in $L^2[0,L]$. Moreover, like the sine-functions in the uniform case, the n^{th} eigenfunction has $(n-1)$ zeros in the interval $0 < x < L$. This is an example of a general oscillation property possessed by Sturm-Liouville eigenfunctions, and it means that eigenfunctions with higher eigenvalues oscillate more rapidly than ones with lower eigenvalues.

2.5. The Helmholtz equation

Analogous results apply in higher space dimensions, although the resulting eigenvalue problems are more difficult to analyze because they involve PDEs instead of ODEs.

Consider the wave equation in n space-dimensions,

$$u_{tt} = c_0^2 \Delta u.$$

We look for time-periodic solutions of the form

$$u(x,t) = \varphi(x) e^{-i\omega t}$$

where the frequency ω is constant. We find that φ satisfies the *Helmholtz* equation

$$-\Delta\varphi = \lambda\varphi, \tag{4.18}$$

where $\lambda = k^2$, and the wavenumber k is given by

$$k = \frac{\omega}{c_0}.$$

Consider solutions of (4.18) on a smooth, bounded domain $\Omega \subset \mathbb{R}^n$, subject to homogeneous Dirichlet boundary conditions

$$\varphi(x) = 0 \qquad \text{for } x \in \partial\Omega \tag{4.19}$$

Equations (4.18)–(4.19) are an eigenvalue problem for the Laplace equation on Ω.

It is possible to show that the eigenvalues form an infinite increasing sequence $0 < \lambda_1 < \lambda_2 \leq \lambda_3 \leq \dots$. If λ_n is an eigenvalue, then $\omega_n = c_0\sqrt{\lambda_n}$ is a resonant frequency of the wave equation. For example, if $n = 2$, we may think of the resonant frequencies of a drum, and if $n = 3$, we may think of the resonant frequencies of sound waves in a container.

Mark Kac [**29**] asked the question: "Can one hear the shape of a drum?" In other words, is it possible to deduce the shape of a planar domain $\Omega \subset \mathbb{R}^2$ given the sequence of Dirichlet eigenvalues $\{\lambda_n\}$ for the Laplacian on Ω.

The sequence of eigenvalues contains a considerable amount of geometrical information. For example, according to Weyl's formula, in n space-dimensions the volume V (or area if $n = 2$) of the domain is given by

$$V = \lim_{R\to\infty} \frac{(2\pi)^n N(R)}{R^{n/2}}$$

where $N(R)$ denotes the number of eigenvalues of the Dirichlet Laplacian that are less than R.

More generally, one can ask a similar question about the whether or not the eigenvalues of the Laplace-Beltrami operator on a Riemannian manifold determine the manifold up to an isometry. Milnor (1964) gave examples of two non-isometric sixteen dimensional tori whose Laplace-Betrami operators have the same eigenvalues. The two-dimensional question remained open until 1992, when Gordon, Webb, and Wolpert constructed two non-isometric plane domains whose Dirichlet Laplacians have the same eigenvalues.

Related inverse spectral problems that involve the reconstruction of the coefficients of a differential operator from appropriate spectral data are important in connection with the theory of completely integrable nonlinear PDEs, such as the KdV equation.

3. Quantum mechanics

> I was in Bristol at the time I started on Heisenberg's theory. I had returned home for the last part of the summer vacation, and I went back to Cambridge at the beginning of October, 1925, and resumed my previous style of life, intensive thinking on the

> problems during the week and relaxing on Sunday, going for a long walk in the country alone. It was during one of these Sunday walks in October, when I was thinking very much about this $uv - vu$, in spite of my intention to relax, that I thought about Poisson brackets.[1]

There is no systematic derivation of quantum mechanics from classical mechanics. (If there were, presumably quantum mechanics would have been discovered by the founders of classical mechanics.) There is, however, a close correspondence between the two theories. One way to understand the correspondence is through path integrals, which leads to the Lagrangian formulation of classical mechanics. Here, we will discuss an alternative, Hamiltonian, approach.

3.1. The correspondence principle

In the Hamiltonian formulation of classical mechanics, observables (such as the energy) are functions defined on the phase space of the system. In the Heisenberg formulation of quantum mechanics, observables are self-adjoint operators acting on a complex Hilbert space. The possible values of a quantum mechanical observable are the elements of its spectrum, and an eigenvector of an observable is a state with a definite value of the observable equal to the associated eigenvalue. The quantum and classical theories correspond in the sense that the commutators of quantum-mechanical operators agree with the canonical Poisson brackets of the corresponding classical functions multiplied by $i\hbar$.

To write this requirement explicitly, we let $\hat{F}$ denote the quantum mechanical operator corresponding to the classical observable F. Then for any pair of classical observables F, G we require that

$$\widehat{\{F,G\}} = \frac{1}{i\hbar}\left[\hat{F},\hat{G}\right].$$

Here, $\{F,G\}$ is the canonical Poisson bracket of F, G, and

$$\left[\hat{F},\hat{G}\right] = \hat{F}\hat{G} - \hat{G}\hat{F}$$

is the commutator of the operators $\hat{F}$, $\hat{G}$. This prescription is dimensionally consistent, since the canonical bracket involves a derivative with respect to momentum and position, which has the dimension of action.

Thus, roughly speaking, the prescription in passing from classical to quantum mechanics is to replace Poisson brackets by commutators divided by $i\hbar$. This prescription is not entirely unambiguous when it leads to products of non-commuting operators, since all such ordering information is lost in the passage from quantum to classical mechanics.

The classical Hamiltonian equations for the evolution of a function F with respect to a Hamiltonian H is

$$F_t = \{F,H\}.$$

Thus, the corresponding quantum mechanical equation is

$$i\hbar\hat{F}_t = \left[\hat{F},\hat{H}\right].$$

[1] P. A. M. Dirac, Varenna lectures, 1972. Dirac apparently had to wait until Monday so he could look up Poisson brackets in the library, such was the speed of transmission of information at the time.

This operator equation has the solution

$$\hat{F}(t) = e^{-i\hat{H}t/\hbar}\hat{F}_0 e^{i\hat{H}t/\hbar}. \tag{4.20}$$

The Hamiltonian $\hat{H}$ is self adjoint, so

$$\left(e^{-i\hat{H}t/\hbar}\right)^* = e^{i\hat{H}^*t/\hbar} = e^{i\hat{H}t/\hbar} = \left(e^{-i\hat{H}t/\hbar}\right)^{-1},$$

meaning that the evolution operator $e^{-i\hat{H}t/\hbar}$ is unitary.

In this 'Heisenberg picture' the operators evolve in time and act on a fixed vector ψ_0 in an underlying Hilbert space $\mathcal{H}$. The measurable quantities associated with an observable $\hat{F}$ are inner products of the form

$$\left\langle \varphi_0, \hat{F}(t)\psi_0 \right\rangle$$

where $\varphi_0, \psi_0 \in \mathcal{H}$, and $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathcal{H}$.

3.2. The Schrödinger equation

To obtain the 'Schrödinger picture,' from the 'Heisenberg picture,' we transfer the time-evolution from the operators to the states. That is, given a fixed vector $\psi_0 \in \mathcal{H}$, we define

$$\psi(t) = e^{i\hat{H}t/\hbar}\psi_0. \tag{4.21}$$

Then, if φ_0, ψ_0 are any two vectors in $\mathcal{H}$, with corresponding time-dependent states $\psi(t)$, $\varphi(t)$, we have from (4.20)

$$\left\langle \varphi_0, \hat{F}(t)\psi_0 \right\rangle = \left\langle \varphi(t), \hat{F}_0\psi(t) \right\rangle.$$

Moreover, since conjugation preserves commutators, the operators $\hat{F}(t)$ and $\hat{F}_0$ satisfy the same commutation relations. Thus, both 'pictures' lead to the same result.

The Schrödinger state vector $\psi(t)$ in (4.21) satisfies the evolution equation

$$i\hbar\psi_t = H\psi.$$

This equation is the Schrödinger equation for a nonrelativistic quantum mechanical system.

Now consider the canonical Hamiltonian formulation of a classical mechanical system with conjugate position and momentum variables,

$$\vec{x} = (x_1, x_2, \ldots, x_n), \qquad \vec{p} = (p_1, p_2, \ldots, p_n).$$

Their Poisson brackets are given by

$$\{x_j, x_k\} = 0, \qquad \{p_j, p_k\} = 0, \qquad \{x_j, p_k\} = \delta_{jk}.$$

We represent these operators as operators acting on functions in $L^2(\mathbb{R}^n)$. We define the position operator $\hat{x}$ as a multiplication operator, and the momentum operator $\hat{p}$ as a gradient:

$$\hat{x} = \vec{x}, \quad \hat{x}_j = x_j; \qquad \hat{p} = -i\hbar\nabla, \quad \hat{p}_k = -i\hbar\frac{\partial}{\partial x_k}.$$

We have

$$[\hat{x}_j, \hat{x}_k] = 0, \qquad [\hat{p}_j, \hat{p}_k] = 0, \qquad [\hat{x}_j, \hat{p}_k] = i\hbar\delta_{jk},$$

in correspondence with the Poisson brackets of the classical position and momentum functions.

We again consider a particle of mass m moving in n-space dimensions in a potential V. of the particles. The kinetic energy operator $T = \hat{p}^2/(2m)$ is given by

$$T = -\frac{\hbar^2}{2m}\Delta,$$

where we now drop the 'hats' on operators. The potential energy operator V is multiplication by $V(\vec{x})$. The Hamiltonian operator $H = T + V$ is therefore

$$H = -\frac{\hbar^2}{2m}\Delta + V(\vec{x}).$$

We describe the state of a quantum-mechanical particle by a complex-valued wavefunction $\psi(\vec{x}, t)$, where $|\psi|^2(\vec{x}, t)$ is the spatial probability density for the location of the particle at time t.

The time-dependent Schrödinger equation in this case is the linear PDE

$$i\hbar\psi_t = -\frac{\hbar^2}{2m}\Delta\psi + V(\vec{x})\psi. \tag{4.22}$$

3.3. Energy eigenfunctions

> In this paper I wish to consider, first, the simple case of the hydrogen atom (no-relativistic and unperturbed), and show that the customary quantum conditions can be replaced by another postulate, in which the notion of "whole numbers," merely as such, is not introduced. Rather, when integrality does appear, it arises in the same natural way as it does in the case of the *node numbers* of a vibrating string. The new conception is capable of generalization, and strikes, I believe, very deeply at the nature of the quantum rules.[2]

Separable solutions of (4.22) of the form

$$\psi(\vec{x}, t) = \varphi(\vec{x})\, e^{-iEt/\hbar}$$

correspond to energy eigenstates with energy E. The function $\varphi(\vec{x})$ satisfies the time-independent Schrödinger equation

$$-\frac{\hbar^2}{2m}\Delta\varphi + V(\vec{x})\varphi = E\varphi.$$

This equation may be supplemented with suitable boundary conditions. For example, if the particle is confined to a bounded domain $\Omega \subset \mathbb{R}^n$, we impose $\varphi = 0$ on $\partial\Omega$ Eigenvalues correspond to the energy levels of bound states, while continuous spectrum corresponds to scattered states.

4. The one-dimensional Schrödinger equation

For a single quantum-mechanical particle of mass m moving in one space dimension in a potential $V(x)$, the time-dependent Shrödinger equation (4.22) is

$$i\hbar\psi_t = -\frac{\hbar^2}{2m}\psi_{xx} + V(x)\psi. \tag{4.23}$$

Looking for separable solutions

$$\psi(x, t) = \varphi(x)\, e^{-iEt/\hbar},$$

[2]E. Schrodinger, translated from *Annalen der Physic***79** (1926).

we find that $\varphi(x)$ satisfies the ODE

$$-\frac{\hbar^2}{2m}\varphi'' + V(x)\,\varphi = E\varphi. \tag{4.24}$$

After normalization, this a Sturm-Liouville equation (4.1) of the form

$$-u'' + qu = \lambda u.$$

The coefficient q is proportional to the potential V and the eigenvalue parameter λ in proportional to the energy E.

4.1. A free particle in a box

Consider a particle that is confined to a finite 'box' of length L, but is otherwise free to move. Formally, this corresponds to a potential

$$V(x) = \begin{cases} 0 & \text{if } 0 < x < L, \\ \infty & \text{otherwise.} \end{cases}$$

Since the probability of finding the particle outside the box $0 \le x \le L$ is zero, continuity of the wavefunction implies that it vanishes at the endpoints, and the spatial profile $\varphi(x)$ of an energy eigenfunction with energy E satisfies the BVP

$$-\frac{\hbar^2}{2m}\varphi'' = E\varphi, \qquad \varphi(0) = 0, \qquad \varphi(L) = 0.$$

This has exactly the same form as the eigenvalue problem arising from the vibration of a uniform string. In particular, the energy levels are

$$E_n = \frac{1}{2m}\left(\frac{n\hbar\pi}{L}\right)^2, \qquad n = 1, 2, 3, \ldots.$$

4.2. Motion in a constant force

Consider a particle that is confined to a half-line $x > 0$ and acted on by a constant force $-F$ directed toward the origin, so that $F > 0$. The corresponding potential is $V(x) = Fx$, and the energy eigenfunction with energy E satisfies the BVP

$$\begin{aligned} -\frac{\hbar^2}{2m}\varphi'' + Fx\varphi &= E\varphi, \\ \varphi(0) &= 0, \\ \varphi(x) &\to 0 \qquad \text{as } x \to \infty. \end{aligned}$$

This problem may be solved in terms of Airy functions $\mathrm{Ai}(x)$, discussed below. Its spectrum consists of eigenvalues, which may be expressed in terms of the zeros of $\mathrm{Ai}(x)$.

Classically, a particle of energy E would repeatedly bounce elastically off the wall at $x = 0$ to a distance $a = E/F$. In quantum mechanics, the wavefunction of the particle is localized near the wall.

4.3. The simple harmonic oscillator

> I remember that when someone had started to teach me about creation and annihilation operators, that this operator creates an electron, I said, "how do you create an electron? It disagrees with the conservation of charge", and in that way, I blocked my mind from learning a very practical scheme of calculation. [3]

[3] Richard P. Feynman, Nobel Lecture, December 11, 1965.

The quadratic potential $V(x) = \frac{1}{2}kx^2$ corresponds to a simple harmonic oscillator. In that case, the energy eigenstates satisfy

$$-\frac{\hbar^2}{2m}\varphi'' + \frac{1}{2}kx^2\varphi = E\varphi. \tag{4.25}$$

We consider this eigenvalue problem on the infinite domain $-\infty < x < \infty$, and look for solutions $\varphi \in L^2(\mathbb{R})$ that decay as $|x| \to \infty$.

Although this problem is posed on the whole real line, its spectrum consists entirely of eigenvalues. This is because it involves an 'oscillator-type' potential, meaning that $V(x) \to +\infty$ as $|x| \to \infty$, so a particle with finite energy is confined to a bounded region with high probability.

Despite the fact that the ODE in (4.25) has variable coefficients, the eigenvalue problem is explicitly solvable in terms of elementary Hermite functions. From the perspective of Feynman path integrals, this is explained by the fact that the corresponding path integral is an oscillatory Gaussian integral, which can be evaluated exactly.

We will solve the problem by the introduction of creation and annihilation, or 'ladder,' operators, which map an eigenfunction to the succeeding, or preceding, eigenfunction. The creation operator adds a quantum of energy to the oscillator, while the annihilation operator removes quantum of energy.

We write (4.25) in operator form as

$$H\varphi = E\varphi \tag{4.26}$$

where the Hamiltonian operator H is given by

$$H = -\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + \frac{1}{2}kx^2. \tag{4.27}$$

We may write H as

$$H = \frac{1}{2m}p^2 + \frac{1}{2}kx^2,$$

where p denotes the momentum operator

$$p = -i\hbar\frac{d}{dx}.$$

Let

$$\omega_0 = \sqrt{\frac{k}{m}}$$

denote the frequency of the corresponding classical simple harmonic oscillator. We define the *annihilation operator* a and the adjoint *creation operator* a^* by

$$a = \sqrt{\frac{k}{2\hbar\omega_0}}\left(x + \frac{ip}{m\omega_0}\right) = \sqrt{\frac{\hbar}{2m\omega_0}}\frac{d}{dx} + \sqrt{\frac{k}{2\hbar\omega_0}}x$$
$$a^* = \sqrt{\frac{k}{2\hbar\omega_0}}\left(x - \frac{ip}{m\omega_0}\right) = -\sqrt{\frac{\hbar}{2m\omega_0}}\frac{d}{dx} + \sqrt{\frac{k}{2\hbar\omega_0}}x,$$

The annihilation and creation operators are dimensionless, and satisfy the commutation relation

$$[a, a^*] = 1.$$

We may express the Hamiltonian in (4.27) in terms of a, a^* as

$$H = \hbar\omega_0\left(aa^* - \frac{1}{2}\right) = \hbar\omega_0\left(a^*a + \frac{1}{2}\right). \tag{4.28}$$

It follows from these equations that

$$[H, a^*] = \hbar\omega_0 a^*, \qquad [H, a] = -\hbar\omega_0 a.$$

Now suppose that φ is an eigenfunction of H, with energy eigenvalue E, so that $H\varphi = E\varphi$. Let $\tilde{\varphi} = a^*\varphi$. Then we find that

$$H\tilde{\varphi} = Ha^*\varphi = a^*H\varphi + [H, a^*]\,\varphi = Ea^*\varphi + \hbar\omega_0 a^*\varphi = \tilde{E}\tilde{\varphi},$$

where $\tilde{E} = E + \hbar\omega_0$. There are no non-zero functions $\varphi \in L^2(\mathbb{R})$ such that $a^*\varphi = 0$, and a^* maps L^2 functions to L^2 functions. Therefore a^* maps an eigenfunction of H with eigenvalue E to one with eigenvalue $E + \hbar\omega_0$.

To start the 'ladder' of eigenfunctions, we observe from (4.28) that if $a\varphi = 0$, then $H\varphi = \frac{1}{2}\hbar\omega_0$. The condition $a\varphi = 0$ corresponds to the ODE

$$\varphi' + \frac{m\omega_0 x}{\hbar}\varphi = 0,$$

which has the L^2-solution

$$\varphi_0(x) = \exp\left(-\frac{m\omega_0 x^2}{2\hbar}\right).$$

It then follows that the n^{th}-eigenfunction φ_n has the form

$$\varphi_n = c_n\,(a^*)^n\,\varphi_0 \tag{4.29}$$

where c_n is any convenient normalization coefficient. The corresponding energy eigenvalues are

$$E_n = \hbar\omega_0\left(n + \frac{1}{2}\right) \qquad \text{where } n = 0, 1, 2, \ldots.$$

The 'ground state' of the oscillator, with $n = 0$, has nonzero energy $\hbar\omega_0/2$. The fact that the energy of the oscillating particle cannot be reduced completely to zero contrasts with the behavior of the corresponding classical oscillator. It may be interpreted as a consequence of the uncertainty principle that the position and momentum of a quantum-mechanical particle cannot both specified simultaneously. As a result, if the particle is located at the point $x = 0$, where the potential energy attains its minimum value of zero, it would necessarily have nonzero momentum and therefore nonzero kinetic energy.

The energy of the n^{th} level is equal to the energy of the ground state plus n 'quanta' of energy $\hbar\omega_0$. This quantization of energy also contrasts with classical mechanics, where the particle can posses any energy $0 \le E < \infty$.

Each derivative in $(a^*)^n$ of the Gaussian φ_0 in (4.29) brings down a factor of x. Thus, the eigenfunctions φ_n have the form of a polynomial of degree n, called a Hermite polynomial, multiplied by the same Gaussian factor. Explicitly, we have

$$\varphi_n(x) = e^{-\alpha^2 x^2/2} H_n(\alpha x), \qquad n = 0, 1, 2, \ldots$$

where the n^{th} Hermite polynomial $H_n(x)$ is defined in (4.31), and

$$\alpha = \sqrt{\frac{m\omega_0}{\hbar}}.$$

The lengthscale α^{-1} is a characteristic lengthscale over which the wavefunction of the ground state of the oscillator varies.

4.4. The Hermite functions

The Hermite functions $\varphi_n(x)$, where $n = 0, 1, 2, \dots$, are eigenfunctions of the Sturm-Liouville equation on $-\infty < x < \infty$

$$-\varphi'' + x^2\varphi = \lambda\varphi. \tag{4.30}$$

The corresponding eigenvalues $\lambda = \lambda_n$ are given by

$$\lambda_n = 2n + 1.$$

Thus, the spectrum σ of (4.30) consists entirely of eigenvalues.

The Hermite functions have the form

$$\varphi_n(x) = e^{-x^2/2} H_n(x),$$

where the Hermite polynomials $H_n(x)$ are given by Rodriguez' formula

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}\left(e^{-x^2}\right). \tag{4.31}$$

We see from this formula that H_n is a polynomial of degree n. Thus, the Hermite functions decay exponentially as $|x| \to \infty$.

First, we show that $\{\varphi_n \mid n = 0, 1, 2, \dots\}$ form an orthogonal set in $L^2(\mathbb{R})$ with respect to the standard inner product

$$\langle f, g\rangle = \int_{-\infty}^{\infty} f(x)g(x)\,dx.$$

It is sufficient to show that φ_n is orthogonal to $e^{-x^2/2}x^m$ for every $0 \le m \le n-1$, since then φ_n is orthogonal to every function of the form $e^{-x^2/2}p_m$ where p_m is a polynomial of degree $m \le n-1$, and hence in particular to φ_m.

Integrating by parts m-times, and using the fact that $e^{-x^2/2}p(x) \to 0$ as $|x| \to \infty$ for every polynomial p, we compute that

$$\begin{aligned}
\left\langle e^{-x^2/2}x^m, \varphi_n \right\rangle &= (-1)^n \int_{-\infty}^{\infty} x^m \frac{d^n}{dx^n}\left(e^{-x^2}\right) dx \\
&= (-1)^{m+n} m! \int_{-\infty}^{\infty} \frac{d^{n-m}}{dx^{n-m}}\left(e^{-x^2}\right) dx \\
&= (-1)^{m+n} m! \left[\frac{d^{n-m-1}}{dx^{n-m-1}}\left(e^{-x^2}\right)\right]_{-\infty}^{\infty} \\
&= 0,
\end{aligned}$$

which proves the result.

Next, we prove that the Hermite polynomials satisfy the following recurrence relations:

$$H_{n+1} = 2xH_n - 2nH_{n-1}, \tag{4.32}$$

$$\frac{dH_n}{dx} = 2nH_{n-1}. \tag{4.33}$$

First, carrying out one differentiation and using the Leibnitz formula for the nth derivative of a product, we get

$$\begin{aligned}
\frac{d^{n+1}}{dx^{n+1}}\left(e^{-x^2}\right) &= \frac{d^n}{dx^n}\left(-2xe^{-x^2}\right) \\
&= -2x\frac{d^n}{dx^n}\left(e^{-x^2}\right) - 2n\frac{d^{n-1}}{dx^{n-1}}\left(e^{-x^2}\right).
\end{aligned}$$

Multiplying this equation by $(-1)^{n+1}e^{x^2}$ and using the definition of the Hermite polynomials, we get (4.32).

Second, using the definition of the Hermite polynomials and the product rule, we get

$$\begin{aligned}\frac{dH_n}{dx} &= (-1)^n \frac{d}{dx}\left[e^{x^2}\frac{d^n}{dx^n}\left(e^{-x^2}\right)\right] \\ &= (-1)^n e^{x^2}\frac{d^{n+1}}{dx^{n+1}}\left(e^{-x^2}\right) + (-1)^n 2x e^{x^2}\frac{d^n}{dx^n}\left(e^{-x^2}\right) \\ &= -H_{n+1} + 2xH_n.\end{aligned}$$

Using (4.32) to eliminate H_{n+1} from this equation, we get (4.33).

To show that the Hermite functions are eigenfunctions of the operator

$$H = -\frac{d^2}{dx^2} + x^2$$

in (4.30), we define annihilation and creation operators

$$a = \frac{d}{dx} + x, \qquad a^* = -\frac{d}{dx} + x, \qquad H = aa^* - 1.$$

Using (4.32)–(4.33), we compute that

$$\begin{aligned}a\varphi_n &= \left(\frac{d}{dx} + x\right)\left(e^{-x^2/2}H_n\right) = e^{-x^2/2}\frac{dH_n}{dx} = 2n\varphi_{n-1}, \\ a^*\varphi_n &= \left(-\frac{d}{dx} + x\right)\left(e^{-x^2/2}H_n\right) = e^{-x^2/2}\left(-\frac{dH_n}{dx} + 2xH_n\right) = \varphi_{n+1}.\end{aligned}$$

It follows that

$$H\varphi_n = (aa^* - 1)\,\varphi_n = a\varphi_{n+1} - \varphi_n = (2n+1)\,\varphi_n,$$

which proves the result.

It is interesting to note that the Hermite functions are eigenfunctions of the Fourier transform. With a convenient normalization, the Fourier transform

$$\mathcal{F} : L^2(\mathbb{R}) \to L^2(\mathbb{R}), \qquad \mathcal{F} : f \mapsto \hat{f}$$

is an isometry on $L^2(\mathbb{R})$. A function f and its Fourier transform $\hat{f}$ are related by

$$f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{f}(k)e^{ikx}\,dk, \qquad \hat{f}(k) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)e^{-ikx}\,dx.$$

Then, using (4.31) in the definition of the Fourier transform, we find that

$$\mathcal{F}[\varphi_n] = (-i)^n\varphi_n.$$

For example, if $n = 0$, this is the familiar fact that the Fourier transform of the Gaussian $e^{-x^2/2}$ is the same Gaussian $e^{-k^2/2}$. As with any unitary map, the spectrum of the Fourier transform lies on the unit circle. It consists of four eigenvalues $1, i, -1, -i$, each of which has infinite multiplicity.

One way to understand why the eigenfunctions of (4.30) are also eigenfunctions of the Fourier transform is to observe that the transforms of a derivative and a product with x are given by

$$\mathcal{F}[f'(x)] = ik\hat{f}(k), \qquad \mathcal{F}[xf(x)] = i\hat{f}'(k).$$

Thus, the operations of differentiation and multiplication by the independent variable exchange places under the Fourier transform. As a result, the operator

$$-\frac{d^2}{dx^2} + x^2 \mapsto k^2 - \frac{d^2}{dk^2}$$

maps to itself. Hence, if φ is an eigenfunction, then $\hat{\varphi}$ is also an eigenfunction.

4.5. Periodic potentials

> When I started to think about it, I felt that the main problem was to explain how the electrons could sneak by all the ions in the metal...By straight Fourier analysis I found to my delight that the wave differed from the plane wave of free electrons only by a periodic modulation.[4]

A simple model for the motion of a conduction electron in a crystal lattice consists of the time-independent Schrödinger equation with a potential that varies periodically in space. This potential describes the effect of the forces due to the ions in the crystal lattice and the other electrons in the crystal on the motion of the electron.

Let us consider the one-dimensional version of this problem. The wavefunction of the electron then satisfies the one-dimensional Schrödinger equation in which the potential $V(x)$ is a periodic function. Suppose that the period is a, so that $V(x+a) = V(x)$. Then, after making the change of variables

$$u(x) = \varphi\left(\frac{x}{a}\right), \qquad q(x) = \frac{2ma^2}{\hbar} V\left(\frac{x}{a}\right), \qquad \lambda = \frac{2mE}{\hbar^2},$$

the normalized wavefunction u and energy parameter λ satisfy

$$-u'' + q\,(x)\,u = \lambda u, \tag{4.34}$$

where $q(x+1) = q(x)$. We will consider (4.34) on the real line $-\infty < x < \infty$ with a general 1-periodic potential q which we assume is continuous.

A specific example of such an equation (with the period of the coefficient q normalized to π rather than 1) is the Mathieu equation

$$-u'' + 2k\cos(2x)u = \lambda u, \tag{4.35}$$

where k is a real parameter. Its solutions, called Mathieu functions, have been studied extensively.

A simpler example to analyze is a 'delta-comb' potential

$$q(x) = Q \sum_{n=-\infty}^{\infty} \delta(x-n),$$

corresponding to a periodic array of δ-potentials at integer lattice points. This equation may be solved explicitly by matching solutions $e^{\pm kx}$ of the 'free' equation with $q = 0$ and $\lambda = k^2$ across the points $x = n$, where u' jumps by Q. Alternatively, one may consider periodic, piecewise constant potentials (the 'Kronig-Penney' model).

As we will see, the spectrum of (4.34) is absolutely continuous, and consists of the union of closed intervals, or 'bands,' separated by 'gaps'. When λ lies inside a band, the equation has bounded, non-square integrable, solutions; when λ lies inside a gap, all nonzero solutions are unbounded, and grow exponentially either at $-\infty$ or ∞.

---

[4] Felix Bloch, quoted in [**31**].

The existence of these bands and gaps has significant implications for the conductivity properties of crystals. Electrons whose energy lies in one of the bands can move freely through the crystal, while electrons whose energy lies in one of the gaps cannot move large distances. A crystal behaves like an insulator if its non-empty energy bands are completely filled with electrons, while it conducts electricity if, like a metal, it has energy bands that are only partially filled (say between 10 and 90 percent). Semiconductors typically have a full valence band, but a small band gap energies E_g of the same order as the thermal energy $k_B T$. As a result, electrons can be thermally excited from the valence band to an empty conduction band. The excited electrons, and the 'holes' they leave behind in the valence band, then conduct electricity (see [**31**] for a detailed discussion).

To study the spectrum of (4.34), we use Floquet theory, which applies to linear ODEs with periodic coefficients. Floquet theory is also used to study the stability of time-periodic solutions of ODEs.

The periodicity of the coefficient $q(x)$ does not imply that solutions are periodic, but it does imply that if $u(x)$ is a solution, then so is $u(x+1)$. For example, the ODE $u'' + u = 0$ trivially has 1-periodic coefficients, since they are constant. The solutions $\cos x$, $\sin x$ are not 1-periodic, but $\sin(x+1) = \sin 1 \cos x + \cos 1 \sin x$ is a solution.

Suppose that, for a given value of λ, the functions $u_1(x;\lambda)$, $u_2(x;\lambda)$ form a fundamental pair of solutions for (4.34). It follows that there are constants $a_{ij}(\lambda)$, $1 \le i, j \le 2$, such that

$$\begin{aligned} u_1(x+1;\lambda) &= a_{11}(\lambda) u_1(x;\lambda) + a_{12}(\lambda) u_2(x;\lambda), \\ u_2(x+1;\lambda) &= a_{21}(\lambda) u_1(x;\lambda) + a_{22}(\lambda) u_2(x;\lambda). \end{aligned}$$

Let

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

If $\rho \in \mathbb{C}$ is an eigenvalue of A with left eigenvector (c_1, c_2) and $v = c_1 u_1 + c_2 u_2$, then it follows that

$$v(x+n;\lambda) = \rho^n(\lambda) v(x,\lambda) \qquad \text{for every } n \in \mathbb{Z}.$$

This solution is bounded if $|\rho| = 1$, otherwise it grows exponentially either as $x \to -\infty$ (if $|\rho| < 1$) or as $x \to +\infty$ (if $|\rho| > 1$). We call the eigenvalues of A *Floquet multipliers*.

The Wronskian $u_1 u_2' - u_2 u_1'$ is a nonzero constant (since its derivative vanishes and u_1, u_2 are independent). It follows that the matrix A has determinant one, so the characteristic equation of A has the form

$$\det(A - \rho I) = \rho^2 - D\rho + 1,$$

where $D(\lambda) = \operatorname{tr} A(\lambda)$. The value of $D(\lambda)$ does not depend in the choice of the fundamental pair of solutions, since the use of another pair leads to a matrix that is similar to A and has the same trace.

If $|D| > 2$, then A has two real eigenvalues ρ_1, ρ_2, one with $|\rho_1| < 1$ and the other with $|\rho_2| > 1$. Thus, the corresponding fundamental solution are unbounded, and (4.34) has no non-zero bounded solutions.

If $|D| < 2$, then A has two complex-conjugate eigenvalues. Since the product of the eigenvalues is equal to 1, the eigenvalues have modulus equal to one, say

$\rho_1 = e^{i\alpha}$, $\rho_2 = e^{-i\alpha}$. We may then write the corresponding pair of fundamental solutions as

$$v_1(x;\lambda) = e^{i\alpha(\lambda)x} p_1(x;\lambda), \qquad v_2(x;\lambda) = e^{-i\alpha(\lambda)x} p_2(x;\lambda),$$

where $p_1(x;\lambda)$, $p_2(x;\lambda)$ are 1-periodic functions of x. These functions, called Bloch waves, are bounded but not square-integrable, so they are not eigenfunctions.

The function $D(\lambda)$ is a continuous function of λ, and the spectrum σ of (4.34) is real and closed. It follows that the spectrum is given by

$$\sigma = \{\lambda \in \mathbb{R} : |D(\lambda)| \leq 2\}.$$

To describe the spectrum in more detail, we need to analyze the function $D(\lambda)$. We will not carry out a general analysis here but we will describe the result (see [**16**] for more information).

As motivation, it is useful to consider the, almost trivial, explicitly solvable case $q(x) = 0$ from the perspective of Floquet theory. In that case (4.34) is

$$-u'' = \lambda u.$$

If $\lambda < 0$, a fundamental pair of solutions of the ODE is

$$u_1(x;\lambda) = e^{-\sqrt{-\lambda}x}, \qquad u_2(x;\lambda) = e^{\sqrt{-\lambda}x}.$$

Thus,

$$u_1(x+1;\lambda) = e^{-\sqrt{-\lambda}} u_1(x;\lambda), \qquad u_2(x+1;\lambda) = e^{\sqrt{-\lambda}} u_2(x;\lambda),$$

and $A(\lambda)$ is a diagonal matrix with trace

$$D(\lambda) = 2\cosh\sqrt{-\lambda}.$$

If $\lambda > 0$, a fundamental pair of solutions of the ODE is

$$u_1(x;\lambda) = \cos\left(\sqrt{\lambda}x\right), \qquad u_2(x;\lambda) = \sin\left(\sqrt{\lambda}x\right),$$

and

$$u_1(x+1;\lambda) = \cos\left(\sqrt{\lambda}\right) u_1(x;\lambda) - \sin\left(\sqrt{\lambda}\right) u_2(x;\lambda),$$
$$u_2(x+1;\lambda) = \sin\left(\sqrt{\lambda}\right) u_1(x;\lambda) + \cos\left(\sqrt{\lambda}\right) u_2(x;\lambda),$$

so

$$D(\lambda) = 2\cos\sqrt{\lambda}.$$

Thus, $|D(\lambda)| \leq 2$ for $0 \leq \lambda < \infty$, corresponding to continuous spectrum. Also note that $D(\lambda) = 2$ at $\lambda = (2m)^2\pi^2$, where the equation has a two-dimensional space of 1-periodic solutions, and $D(\lambda) = -2$ at $\lambda = (2m+1)^2\pi^2$, where the equation has a two-dimensional space of 2-periodic solutions.

For nonzero periodic potentials, the behavior of $D(\lambda)$ is similar, except that its local maximum values are, in general, greater than 2, and its local minimum values, are in general, less than -2. This leads to a structure with bands of continuous spectrum separated by gaps.

Specifically, given a periodic potential $q(x)$, we introduce two auxiliary eigenvalue problems. The first eigenvalue problem on $0 \leq x \leq 1$ is for 1-periodic solutions of (4.34),

$$-u'' + q(x)\,u = \lambda u,$$
$$u(0) = u(1), \qquad u'(0) = u'(1).$$

This is a regular Sturm-Liouville eigenvalue problem, and its spectrum consists of an infinite sequence of real eigenvalues

$$\lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \dots$$

such that $\lambda_n \to \infty$ as $n \to \infty$. Here, if there are any eigenvalues of multiplicity two (meaning that it has two linearly independent eigenfunctions), we include them twice in the sequence.

The second the eigenvalue problem on $0 \leq x \leq 1$ is for semi-periodic solutions of (4.34), which satisfy

$$\begin{aligned} -u'' + q(x)\,u &= \mu u, \\ u(0) = -u(1), &\qquad u'(0) = -u'(1). \end{aligned}$$

The spectrum of this eigenvalue problem also consists of an infinite sequence of real eigenvalues

$$\mu_0 \leq \mu_1 \leq \mu_2 \leq \dots$$

such that $\mu_n \to \infty$ as $n \to \infty$, where again any eigenvalue of multiplicity two appears twice. The corresponding eigenfunctions extend to 2-periodic functions on $\mathbb{R}$.

One can prove that [**16**]:

(a) $\lambda_0 < \mu_0 \leq \mu_1 < \lambda_1 \leq \lambda_2 < \mu_2 \leq \mu_3 < \lambda_3 \leq \lambda_4 < \dots$;

(b) $D(\lambda)$ decreases from 2 to -2 in the intervals $[\lambda_{2m}, \mu_{2m}]$;

(c) $D(\lambda)$ increases from -2 to 2 in the intervals $[\mu_{2m+1}, \lambda_{2m+1}]$;

(d) $D(\lambda) > 2$ in the intervals $(-\infty, \lambda_0)$ and $(\lambda_{2m+1}, \lambda_{2m+2})$;

(e) $D(\lambda) < -2$ in the intervals (μ_{2m}, μ_{2m+1}).

Thus, the spectrum σ of (4.34) is given by

$$\sigma = \bigcup_{m=0}^{\infty} [\lambda_{2m}, \mu_{2m}] \cup [\mu_{2m+1}, \lambda_{2m+1}]\,.$$

It is purely absolutely continuous, and consists of an infinite sequence of 'bands,' or 'stability intervals,' separated by 'gaps,' or 'instability intervals,' $(\lambda_{2m+1}, \lambda_{2m+2})$ or (μ_{2m}, μ_{2m+1}).

In the case when $\lambda_{2m+1} = \lambda_{2m+2}$, or $\mu_{2m} = \mu_{2m+1}$, is a double eigenvalue of the auxiliary eigenvalue problem, the corresponding 'gap' disappears. For instance, all of the gaps disappear if $q = 0$. On the other hand, for the Mathieu equation (4.35) with $k \neq 0$, every gap has nonzero width, which tends to zero rapidly as $m \to \infty$.

An interesting special class of potentials are the 'finite-gap' potentials, in which $\lambda_{2m+1} = \lambda_{2m+2}$ and $\mu_{2m} = \mu_{2m+1}$ for all but finitely many m. An example of an n-gap potentials is the Lamé equation

$$-u'' + n(n+1)\wp(x)u = \lambda u$$

where the elliptic function $\wp$ is the Weierstrass 'p'-function. These results are of interest in the theory of completely integrable systems, in connection with the use of the inverse scattering transform for spatially-periodic solutions the KdV equation.

Generalizations of these results apply to the time-independent Schrödinger equation with a periodic potential in higher space dimensions, including the existence of Bloch waves. The analysis there is more complicated, and there are

many more possible lattice symmetries in two and three space dimensions than the single periodic lattice in one space dimension.

4.6. Anderson localization

The spectrum of the one-dimensional, time-independent Schrödinger equation on the real line with a continuous periodic potential is always absolutely continuous. For values of λ in the spectrum, the ODE has bounded solutions which do not decay at infinity.

Anderson (1958) observed that random stationary potentials, such as ones that model a disordered medium, can have a dense point spectrum with associated, exponentially decaying eigenfunctions. This phenomenon is called *Anderson localization.*

As an example, consider a Schrödinger equation of the form

$$-u'' + \left(\sum_{n \in \mathbb{Z}} Q_n(\omega) f(x-n) \right) u = \lambda u \qquad -\infty < x < \infty, \tag{4.36}$$

where $f(x)$ is a given potential function, which is the same at different lattice points n, and the amplitudes $Q_n(\omega)$ are independent identically distributed random variables.

Then, under suitable assumptions (for example, $f(x) \geq 0$ has support in $[0,1]$, so the supports of f do not overlap, and the Q_n are independent random variables uniformly distributed on $[0,1]$) the spectrum of (4.36) is, almost surely, the interval $0 \leq \lambda < \infty$. Moreover, it is pure point spectrum, meaning that there are countably many eigenvalues which are dense in $[0,\infty)$ (similar to the way in which the rational numbers are dense in the real numbers).

Localization has also been studied and observed in classical waves, such as waves in an elastic medium.

5. The Airy equation

The eigenvalue equation for Airy's equation is

$$-u'' + xu = \lambda u.$$

In this case, we can remove the spectral parameter λ by a translation $x \mapsto x - \lambda$, so we set $\lambda = 0$ to get

$$-u'' + xu = 0. \tag{4.37}$$

(The Airy operator on the real line has continuous spectrum $\mathbb{R}$, with bounded solutions given by translations of the Airy function described below.)

The coefficient of the lower order term in (4.37) changes sign at $x = 0$. As a result, one might expect that the qualitative behavior of its solutions changes from oscillatory (like $u'' + u = 0$) when x is large and negative to exponential (like $u'' - u = 0$) when x is large and positive. This is indeed the case, and the Airy functions are, perhaps, the most fundamental functions that describe a continuous transition from oscillatory to exponential behavior as a real variable changes.

One of the most familiar example of this phenomenon occurs at the bright caustics one can observe in light reflections. Airy functions describe the high-frequency light wave-field near a smooth convex caustic that separates the illuminated region from the shadow region. Similar problems arise in semi-classical quantum mechanics, where the wavefunction of a particle is oscillatory in classically allowed regions,

and exponential in classically forbidden regions. The Airy functions describe the transition between these two regimes. The fact that the Airy functions have an exponentially decaying tail is what allows a quantum mechanical particle to 'tunnel' through a classically impassible potential barrier. Here, we will describe an application of Airy functions to the propagation of linear dispersive waves.

First, we summarize some properties of the Airy functions. A standard fundamental pair of solutions of (4.37) is denoted by $\mathrm{Ai}(x)$ and $\mathrm{Bi}(x)$. The solution Ai is determined uniquely, up to a normalization constant, by the condition that it decays exponentially as $x \to \infty$. The function Bi is a second, independent solution of (4.37), which grows exponentially as $x \to \infty$. This property does not determine Bi up to normalization, since we could add to it any multiple of Ai without altering this asymptotic behavior.

These solutions may be defined by their initial values at $x = 0$:

$$\mathrm{Ai}(0) = \alpha, \quad \mathrm{Ai}'(0) = -\beta, \qquad \mathrm{Bi}(0) = \sqrt{3}\alpha, \quad \mathrm{Bi}'(0) = \sqrt{3}\beta.$$

Here, the constants $\alpha \approx 0.355$, $\beta \approx 0.259$ are defined by

$$\alpha = \frac{1}{3^{2/3}\Gamma(2/3)}, \qquad \beta = \frac{1}{3^{1/3}\Gamma(1/3)}$$

where the Gamma-function Γ is defined by

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt, \qquad \text{for } x > 0,$$

An integration by parts shows that $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$, so the Gamma-function may be regarded as an extension of the factorial to non-integers.

In order to properly understand the behavior of the Airy functions it is necessary to consider them in the complex plane. For $z \in \mathbb{C}$, using a Fourier-Laplace transform, we write

$$u(z) = \int_C e^{z\xi} f(\xi)\, d\xi \tag{4.38}$$

where C is a suitable contour and $f : C \to \mathbb{C}$ is a function.

Then, assuming we can differentiate under the integral sign and integrate by parts, we find that

$$-\frac{d^2u}{dz^2} + zu = \int_C e^{z\xi} \left(\xi^2 f + \frac{df}{d\xi} \right) d\xi.$$

Thus, $u(z)$ is a solution of the Airy equation if $f(\xi)$ satisfies

$$f' + \xi^2 f = 0.$$

The simplification here is that, since $u(z)$ is multiplied by the first power of z in the original equation, we get a first-order ODE for $f(\xi)$, which is easy to solve. Up to a constant factor, the solution is

$$f(\xi) = e^{-\xi^3/3}.$$

Suppose that the contour C is given parametrically as $\xi = \xi(t)$ with $-\infty < t < \infty$. In order to ensure convergence of the contour integral in (4.38), we require that $\xi(t) \sim te^{2\pi ik/3}$ for some $k = -1, 0, 1$ as $t \to \infty$, in which case $\xi^3(t) \sim t^3$, and $\xi(t) \sim -te^{2\pi ik/3}$ for some $k = -1, 0, 1$ as $t \to -\infty$, in which case $\xi^3(t) \sim -t^3$.

Up to orientation, this gives three types of contours C_1, C_2, C_3. We define

$$E_k(z) = \frac{1}{\pi i} \int_{C_k} e^{z\xi - \xi^3/3}\, d\xi.$$

Then, since the integrand has no singularities in $\mathbb{C}$ and $C_1 + C_2 + C_3$ is a closed curve (after being deformed away from infinity), Cauchy's theorem implies that

$$E_1(z) + E_2(z) + E_3(z) = 0$$

Also

$$E_2(z) = -\overline{E_2}(z).$$

One can show that the Airy functions are defined so that

$$E_3(z) = \mathrm{Ai}(z) + i\,\mathrm{Bi}(z).$$

These functions are entire (that is, analytic in all of $\mathbb{C}$), with an essential singularity at ∞.

Deforming the contours C_3 to the real axis, we may derive a Fourier representation of the Airy functions as oscillatory integrals

$$\begin{aligned}
\mathrm{Ai}(x) &= \frac{1}{\pi} \int_0^\infty \cos\left(\frac{1}{3}t^3 + xt\right)\, dt, \\
\mathrm{Bi}(x) &= \frac{1}{\pi} \int_0^\infty \left[e^{-t^3/3+xt} + \sin\left(\frac{1}{3}t^3 + xt\right)\right]\, dt
\end{aligned}$$

Note that, in comparison with the Fresnel integral, the oscillatory integrals for the Airy functions have two stationary phase points at $t = \pm\sqrt{-x}$ when $x < 0$ and no stationary phase points when $x > 0$. This explains their transition from oscillatory to exponential behavior.

Using the method of steepest descent, one can show that the Airy functions have the asymptotic behaviors

$$\begin{aligned}
\mathrm{Ai}(x) &\sim \frac{\sin\left(2|x|^{3/2}/3 + \pi/4\right)}{\sqrt{\pi}|x|^{1/4}} && \text{as } x \to -\infty, \\
\mathrm{Bi}(x) &\sim \frac{\cos\left(2|x|^{3/2}/3 + \pi/4\right)}{\sqrt{\pi}|x|^{1/4}} && \text{as } x \to -\infty, \\
\mathrm{Ai}(x) &\sim \frac{\exp\left(-2x^{3/2}/2\right)}{2\sqrt{\pi}x^{1/4}} && \text{as } x \to \infty, \\
\mathrm{Bi}(x) &\sim \frac{\exp\left(2x^{3/2}/3\right)}{2\sqrt{\pi}x^{1/4}} && \text{as } x \to \infty.
\end{aligned}$$

6. Dispersive wave propagation

An important application of the method of stationary phase, discussed briefly in Section 14.2, concerns the long-time, or large-distance, behavior of linear dispersive waves. Kelvin (1887) originally developed the method for this purpose, following earlier work by Cauchy, Stokes, and Riemann. He used it to study the pattern of dispersive water waves generated by a ship in steady motion, and showed that at large distances from the ship the waves form a wedge with a half-angle of $\sin^{-1}(1/3)$, or approximately 19.5°.

Here, we will illustrate this method by using it to study the linearized Korteweg-de Vries (KdV), or Airy equation. We will then show how Airy functions arise

when two stationary phase point coalesce, leading to a transition from oscillatory to exponential behavior.

Consider the following initial value problem (IVP) for the linearized KdV equation

$$\begin{aligned} u_t &= u_{xxx}, \\ u(x,0) &= f(x). \end{aligned}$$

This equation provides an asymptotic description of linear, unidirectional, weakly dispersive long waves. It was first derived for shallow water waves. In the following section we give a derivation of the KdV equation for ion-acoustic waves in a plasma.

We assume for simplicity that the initial data $f : \mathbb{R} \to \mathbb{R}$ is a Schwarz function, meaning that it is smooth and decays, together with all its derivatives, faster than any polynomial as $|x| \to \infty$.

Let $\widehat{u}(k,t)$ denote the Fourier transform of $u(x,t)$ with respect to x,

$$\begin{aligned} u(x,t) &= \int_{-\infty}^{\infty} \widehat{u}(k,t) e^{ikx}\, dk, \\ \widehat{u}(k,t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} u(x,t) e^{-ikx}\, dx. \end{aligned}$$

Then $\widehat{u}(k,t)$ satisfies

$$\begin{aligned} \widehat{u}_t + ik^3 \widehat{u} &= 0, \\ \widehat{u}(k,0) &= \widehat{f}(k). \end{aligned}$$

The solution of this equation is

$$\widehat{u}(k,t) = \widehat{f}(k) e^{-i\omega(k)t},$$

where

$$\omega(k) = k^3.$$

The function $\omega : \mathbb{R} \to \mathbb{R}$ gives the (angular) frequency $\omega(k)$ of a wave with wavenumber k, and is called the *dispersion relation* of the KdV equation.

Inverting the Fourier transform, we find that the solution is given by

$$u(x,t) = \int_{-\infty}^{\infty} \widehat{f}(k) e^{ikx - i\omega(k)t}\, dk.$$

Using the convolution theorem, we can write this solution as

$$u(x,t) = f * g(x,t), \tag{4.39}$$

where the star denotes convolution with respect to x, and

$$g(x,t) = \frac{1}{(3t)^{1/3}} \operatorname{Ai}\left(-\frac{x}{(3t)^{1/3}} \right)$$

is the Green's function of the Airy equation.

This Green's function may also be found directly by looking for similarity solutions

$$g(x,t) = \frac{1}{t^m} G\left(\frac{x}{t^n} \right)$$

of the linearized KdV equation such that

$$\int_{-\infty}^{\infty} g(x,t)\, dx \to 1 \qquad \text{as } t \to 0.$$

We consider the asymptotic behavior of the solution (4.39) as $t \to \infty$ with $x/t = v$ fixed. This limit corresponds to the large-time limit in a reference frame moving with velocity v.

Thus, we want to find the behavior as $t \to \infty$ of

$$u(vt,t) = \int_{-\infty}^{\infty} \widehat{f}(k) e^{i\varphi(k,v)t}\, dk, \tag{4.40}$$

where

$$\varphi(k,v) = kv - \omega(k).$$

The stationary phase points satisfy $\varphi_k = 0$, or

$$v = \omega'(k).$$

The solutions are the wavenumbers k whose group velocity $\omega'(k)$ is equal to v. It follows that

$$3k^2 = v.$$

If $v < 0$, then there are no stationary phase points, and $u(vt,t) = o(t^{-n})$ as $t \to \infty$ for any $n \in \mathbb{N}$.

If $v > 0$, then there are two nondegenerate stationary phase points at $k = \pm k_0(v)$, where

$$k_0(v) = \sqrt{\frac{v}{3}}.$$

These two points contribute complex conjugate terms, and the method of stationary phase implies that

$$u(vt,t) \sim \sqrt{\frac{2\pi}{|\omega''(k_0)|t}}\, \widehat{f}(k_0) e^{i\varphi(k_0,v)t - i\pi/4} + \text{c.c.} \qquad \text{as } t \to \infty.$$

The energy in the wave-packet therefore propagates at the group velocity $C = \omega'(k)$,

$$C = 3k^2,$$

rather than the phase velocity $c = \omega/k$,

$$c = k^2.$$

The solution decays at a rate of $t^{-1/2}$, in accordance with the linear growth in t of the length of the wavetrain and the conservation of energy,

$$\int_{-\infty}^{\infty} u^2(x,t)\, dt = \text{constant}.$$

The two stationary phase points coalesce when $v = 0$, and then there is a single degenerate stationary phase point. To find the asymptotic behavior of the solution when v is small, we make the change of variables

$$k = \frac{\xi}{(3t)^{1/3}}$$

in the Fourier integral solution (4.40). This gives

$$u(x,t) = \frac{1}{(3t)^{1/3}} \int_{-\infty}^{\infty} \widehat{f}\left(\frac{\xi}{(3t)^{1/3}}\right) e^{-i\left(\xi w + \frac{1}{3}\xi^3\right)}\, d\xi,$$

where

$$w = -\frac{t^{2/3} v}{3^{1/3}}.$$

It follows that as $t \to \infty$ with $t^{2/3}v$ fixed,

$$u(x,t) \sim \frac{2\pi}{(3t)^{1/3}} \widehat{f}(0)\, \mathrm{Ai}\left(-\frac{t^{2/3}v}{3^{1/3}}\right).$$

Thus the transition between oscillatory and exponential behavior is described by an Airy function. Since $v = x/t$, the width of the transition layer is of the order $t^{1/3}$ in x, and the solution in this region is of the order $t^{-1/3}$. Thus it decays more slowly and is larger than the solution elsewhere.

7. Derivation of the KdV equation for ion-acoustic waves

The Korteweg-de Vries (KdV) equation is the following nonlinear PDE for $u(x,t)$:

$$u_t + uu_x + u_{xxx} = 0. \tag{4.41}$$

This equation was first derived by Korteweg and de Vries (1895) for shallow water waves, and it is a generic asymptotic equation that describes weakly nonlinear waves with weak long-wave dispersion.

The term u_t is the rate of change of the wave profile u in a reference frame moving with the linearized phase velocity of the wave. The term uu_x is an advective nonlinearity, and u_{xxx} is a linear dispersive term.

Water waves are described by a relatively complicated system of equations which involve a free boundary. Here, we derive the KdV equation from a simpler system of PDEs that describes ion acoustic waves in a plasma. This derivation illustrates the universal nature of the KdV equation, which applies to any wave motion with weak advective nonlinearity and weak long wave dispersion. Specifically, the linearized dispersion relation $\omega = \omega(k)$ between frequency ω and wavenumber k should have a Taylor expansion as $k \to 0$ of the form $\omega = c_0 k + \alpha k^3 + \dots$.

7.1. Plasmas

A plasma is an ionized fluid consisting of positively charged ions and negatively charged electrons which interact through the electro-magnetic field they generate. Plasmas support waves analogous to sound waves in a simple compressible fluid, but as a result of the existence of ion and electron oscillations in plasmas, these waves are dispersive.

Here, we use a simple 'two-fluid' model of a plasma in which the ions and electrons are treated as separate fluids. More detailed models use a kinetic description of the plasma. The full system of equations follows from the fluid equations for the motion of the ions and electrons, and Maxwell's equations for the electro-magnetic field generated by the charged fluids. We will consider relatively low frequency waves that involve the motion of the ions, and we assume that there are no magnetic fields. After simplification and nondimensionalization, we get the equations summarized in (4.47) below.

Let n^i, n^e denote the number density of the ions and electrons, respectively, u^i, u^e their velocities, p^i, p^e their pressures, and E the electric field.

In one space dimension, the equations of conservation of mass and momentum for the ion fluid are

$$\begin{aligned} &n^i_t + (n^i u^i)_x = 0, \\ &m^i n^i \left(u^i_t + u^i u_{ix}\right) + p^i_x = e n^i E. \end{aligned}$$

Here, m^i is the mass of an ion and e is its charge. For simplicity, we assume that this is the same as the charge of an electron.

We suppose that the ion-fluid is 'cold', meaning that we neglect its pressure. Setting $p^i = 0$, we get

$$\begin{aligned} &n^i_t + (n^i u^i)_x = 0, \\ &m^i \left(u^i_t + u^i u_{ix}\right) = eE. \end{aligned}$$

The equations of conservation of mass and momentum for the electron fluid are

$$\begin{aligned} &n^e_t + (n^e u^e)_x = 0, \\ &m^e n^e \left(u^e_t + u^e u^e_x\right) + p^e_x = -e n^e E, \end{aligned}$$

where m^e is the mass of an electron and $-e$ is its charge. The electrons are much lighter than the ions, so we neglect their inertia. Setting $m^e = 0$, we get

$$p^e_x = -e n^e E. \tag{4.42}$$

As we will see, this equation provides an equation for the electron density n^e. The electron velocity u^e is then determined from the equation of conservation of mass. It is uncoupled from the remaining variables, so we do not need to consider it further.

We assume an isothermal equation of state for the electron fluid, meaning that

$$p^e = kTn^e, \tag{4.43}$$

where k is Boltzmann's constant and T is the temperature. Using (4.43) in (4.42) and writing $E = -\varphi_x$ in terms of an electrostatic potential φ, we get

$$kTn^e_x = en^e\varphi_x.$$

This equation implies that n^e is given in terms of φ by

$$n^e = n_0 \exp\left(\frac{e\varphi}{kT}\right), \tag{4.44}$$

where the constant n_0 is the electron number density at $\varphi = 0$.

Maxwell's equation for the electrostatic field E generated by a charge density σ is $\epsilon_0 \nabla \cdot E = \sigma$, where ϵ_0 is a dielectric constant. This equation implies that

$$\epsilon_0 E_x = e(n^i - n^e). \tag{4.45}$$

In terms of the potential φ, equation (4.45) becomes

$$-\varphi_{xx} = \frac{e}{\epsilon_0}\left(n^i - n^e\right). \tag{4.46}$$

We may then use (4.44) to eliminate n^e from (4.46).

Dropping the i-superscript on the ion-variables (n^i, u^i), we may write the final system of equations for (n, u, φ) as

$$\begin{aligned} &n_t + (nu)_x = 0, \\ &u_t + uu_x + \frac{e}{m}\varphi_x = 0, \\ &-\varphi_{xx} + \frac{en_0}{\epsilon_0}\exp\left(\frac{e\varphi}{kT}\right) = \frac{e}{\epsilon_0}n. \end{aligned}$$

This system consists of a pair of evolution equations for (n, u) coupled with a semi-linear elliptic equation for φ.

To nondimensionalize these equations, we introduce the the Debye length λ_0 and the ion-acoustic sound speed c_0, defined by

$$\lambda_0^2 = \frac{\epsilon_0 kT}{n_0 e^2}, \qquad c_0^2 = \frac{kT}{m}.$$

These parameters vary by orders of magnitudes for plasmas in different conditions. For example, a dense laboratory plasma may have $n_0 \approx 10^{20}\,\mathrm{m}^{-3}$, $T \approx 60,000\,\mathrm{K}$ and $\lambda_0 \approx 10^{-6}\,\mathrm{m}$; the solar wind near the earth has $n_0 \approx 10^{7}\,\mathrm{m}^{-3}$, $T \approx 120,000\,\mathrm{K}$, and $\lambda_0 \approx 10\,\mathrm{m}$.

Introducing dimensionless variables

$$\bar{x} = \frac{x}{\lambda_0}, \quad \bar{t} = \frac{c_0 t}{\lambda_0}, \quad \bar{n} = \frac{n}{n_0}, \quad \bar{u} = \frac{u}{c_0}, \quad \bar{\varphi} = \frac{e\varphi}{kT},$$

and dropping the 'bars', we get the nondimensionalized equations

$$\begin{aligned} &n_t + (nu)_x = 0, \\ &u_t + uu_x + \varphi_x = 0, \\ &-\varphi_{xx} + e^{\varphi} = n. \end{aligned} \tag{4.47}$$

7.2. Linearized equations

First, we derive the linearized dispersion relation of ion acoustic waves. Linearizing the system (4.47) about $n = 1$, $\varphi = 0$ and $u = 0$, we get

$$\begin{aligned} &n_t + u_x = 0, \\ &u_t + \varphi_x = 0, \\ &-\varphi_{xx} + \varphi = n, \end{aligned}$$

where n now denotes the perturbation in the number density about 1.

We seek Fourier solutions

$$n(x,t) = \hat{n} e^{ikx - i\omega t}, \quad u(x,t) = \hat{u} e^{ikx - i\omega t}, \quad \varphi(x,t) = \hat{\varphi} e^{ikx - i\omega t}.$$

From the last equation, we find that

$$\hat{\varphi} = \frac{\hat{n}}{1 + k^2}.$$

From the first and second equations, after eliminating $\hat{\varphi}$, we get

$$\begin{pmatrix} -i\omega & ik \\ ik/(1+k^2) & -i\omega \end{pmatrix} \begin{pmatrix} \hat{n} \\ \hat{u} \end{pmatrix} = 0$$

This linear system has a non-zero solution if the determinant of the matrix is zero, which implies that (ω, k) satisfies the dispersion relation

$$\omega^2 = \frac{k^2}{1 + k^2}.$$

The corresponding null-vector is

$$\begin{pmatrix} \hat{n} \\ \hat{u} \end{pmatrix} = \hat{a} \begin{pmatrix} k \\ \omega \end{pmatrix},$$

where $\hat{a}$ is an arbitrary constant.

The phase velocity $c = \omega/k$ of these waves is given

$$c = \frac{1}{(1 + k^2)^{1/2}},$$

so that $c \to 1$ as $k \to 0$ and $c \to 0$ as $k \to \infty$.

The group velocity $C = d\omega/dk$ is given by

$$C = \frac{1}{(1+k^2)^{3/2}}.$$

For these waves, the group veclocity is smaller than the phase velocity for all $k > 0$.

In the long-wave limit $k \to 0$, we get the leading order approximation $\omega = k$, corresponding to non-dispersive sound waves with phase speed $\omega/k = 1$. In the original dimensional variables, this speed is the ion-acoustic speed c_0, and the condition for long-wave dispersion to be weak is that $k\lambda_0 \ll 1$, meaning that the wavelength is much larger than the Debye length. In these long waves, the electrons oscillate with the ions, and the fluid behaves essentially like a single fluid. The inertia of the wave is provided by the ions and the restoring pressure force by the electrons.

By contrast, in the short-wave limit $k \to \infty$, we get waves with constant frequency $\omega = 1$, corresponding in dimensional terms to the ion plasma frequency $\omega_0 = c_0/\lambda_0$. In these short waves, the ions oscillate in an essentially fixed background of electrons.

For water waves, the condition for weak long-wave dispersion is that the wavelength is much larger than the depth of the fluid. Such waves are called 'shallow water waves.'

At the next order in k, we find that

$$\omega = k - \frac{1}{2}k^3 + O(k^5) \qquad \text{as } k \to 0. \tag{4.48}$$

The $O(k^3)$ correction corresponds to weak KdV-type long-wave dispersion.

For very long waves, we may neglect φ_{xx} in comparison with e^{φ} in (4.47), which gives $n = e^{\varphi}$ and $n_x = n\varphi_x$. In that case, (n, u) satisfy the isothermal compressible Euler equations

$$\begin{aligned} &n_t + (nu)_x = 0, \\ &n\,(u_t + uu_x) + n_x = 0. \end{aligned}$$

These equations form a nondispersive hyperbolic system. (The analogous system for water waves is the shallow water equations.) In general, solutions form shocks, but then the long-wave approximation breaks down and it is no longer self-consistent.

A weakly nonlinear expansion of these long wave equations, which is a limiting case of the KdV expansion given below,

$$\begin{pmatrix} n \\ u \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \epsilon a(x - t, \epsilon t) \begin{pmatrix} 1 \\ 1 \end{pmatrix} + O(\epsilon^2),$$

leads to an inviscid Burgers equation for $a(\xi, \tau)$,

$$a_\tau + aa_\xi = 0.$$

In the next section, we apply a similar expansion to (4.47) and include the effect of weak long wave dispersion, leading to a KdV equation.

7.3. KdV expansion

We can see from the KdV equation (4.41) what orders of magnitude of the wave amplitude and the spatial and temporal scales lead to a balance between weak nonlinearity and long-wave dispersion. We need u to have the same order of magnitude

as ∂_x^2 and ∂_t to have the same order of magnitude as ∂_x^3. Thus, we want

$$u = O(\epsilon), \qquad \partial_x = O(\epsilon^{1/2}), \qquad \partial_t = O(\epsilon^{3/2})$$

where ϵ is a small positive parameter. We could, of course, replace ϵ by ϵ^2, or some other small parameter, provided that we retain the same relative scalings. Here, the time-derivative ∂_t is taken in a reference frame moving with the linearized wave velocity.

This scaling argument suggests that we seek an asymptotic solution of (4.47), depending on a small parameter ϵ of the form

$$\begin{aligned} n &= n\left(\epsilon^{1/2}(x-\lambda t), \epsilon^{3/2}t; \epsilon\right), \\ u &= u\left(\epsilon^{1/2}(x-\lambda t), \epsilon^{3/2}t; \epsilon\right), \\ \varphi &= \varphi\left(\epsilon^{1/2}(x-\lambda t), \epsilon^{3/2}t; \epsilon\right). \end{aligned}$$

We will determine the wave velocity λ as part of the solution. The parameter ϵ does not appear explicitly in the PDE (4.47), but it could appear in the initial conditions, for example.

We introduce multiple-scale variables

$$\xi = \epsilon^{1/2}(x-\lambda t), \qquad \tau = \epsilon^{3/2}t.$$

According to the chain rule, we may expand the original space-time derivatives as

$$\partial_x = \epsilon^{1/2}\partial_\xi, \qquad \partial_t = -\epsilon^{1/2}\lambda\partial_\xi + \epsilon^{3/2}\partial_\tau.$$

After including the small parameter ϵ explicitly in the new variables, we assume that derivatives with respect to ξ, τ are of the order 1 as $\epsilon \to 0^+$, which is not the case for derivatives with respect to the original variables x, t.

It follows that $n(\xi,\tau;\epsilon)$, $u(\xi,\tau;\epsilon)$, $\varphi(\xi,\tau;\epsilon)$ satisfy

$$\begin{aligned} &(nu)_\xi - \lambda n_\xi + \epsilon n_\tau = 0, \\ &\varphi_\xi - \lambda u_\xi + u u_\xi + \epsilon u_\tau = 0, \\ &e^\varphi - \epsilon\varphi_{\xi\xi} = n. \end{aligned} \tag{4.49}$$

We look for an asymptotic solution of (4.49) of the form

$$\begin{aligned} n &= 1 + \epsilon n_1 + \epsilon^2 n_2 + \epsilon^3 n_3 + O(\epsilon^4), \\ u &= \epsilon u_1 + \epsilon^2 u_2 + \epsilon^3 u_3 + O(\epsilon^4), \\ \varphi &= \epsilon\varphi_1 + \epsilon^2\varphi_2 + \epsilon^3\varphi_3 + O(\epsilon^4). \end{aligned}$$

Using these expansions in (4.49), Taylor expanding the result with respect to ϵ, and equating coefficients of ϵ, we find that

$$\begin{aligned} &u_{1\xi} - \lambda n_{1\xi} = 0, \\ &\varphi_{1\xi} - \lambda u_{1\xi} = 0, \\ &\varphi_1 - n_1 = 0. \end{aligned} \tag{4.50}$$

Equating coefficients of ϵ^2, we find that

$$\begin{aligned} &u_{2\xi} - \lambda n_{2\xi} + n_{1\tau} + (n_1 u_1)_\xi = 0, \\ &\varphi_{2\xi} - \lambda u_{2\xi} + u_{1\tau} + u_1 u_{1\xi} = 0, \\ &\varphi_2 - n_2 + \frac{1}{2}\varphi_1^2 - \varphi_{1\xi\xi} = 0. \end{aligned} \tag{4.51}$$

Eliminating φ_1 from (4.50), we get a homogeneous linear system for (n_1, u_1),

$$\begin{pmatrix} -\lambda & 1 \\ 1 & -\lambda \end{pmatrix} \begin{pmatrix} n_1 \\ u_1 \end{pmatrix}_\xi = 0.$$

This system has a nontrivial solution if $\lambda^2 = 1$. We suppose that $\lambda = 1$ for definiteness, corresponding to a right-moving wave. Then

$$\begin{pmatrix} n_1 \\ u_1 \end{pmatrix} = a(\xi,\tau) \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \varphi_1 = a(\xi,\tau), \tag{4.52}$$

where $a(\xi, \tau)$ is an arbitrary scalar-valued function.

At the next order, after setting $\lambda = 1$ and eliminating φ_2 in (4.51), we obtain a nonhomogeneous linear system for (n_2, u_2),

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} n_2 \\ u_2 \end{pmatrix}_\xi + \begin{pmatrix} n_{1\tau} + (n_1 u_1)_\xi \\ u_{1\tau} + u_1 u_{1\xi} - \varphi_1\varphi_{1\xi} + \varphi_{1\xi\xi\xi} \end{pmatrix} = 0. \tag{4.53}$$

This system is solvable for (n_2, u_2) if and only if the nonhomogeneous term is orthogonal to the null-vector $(1, 1)$. Using (4.52), we find that this condition implies that $a(\xi, \tau)$ satisfies a KdV equation

$$a_\tau + a a_\xi + \frac{1}{2} a_{\xi\xi\xi} = 0. \tag{4.54}$$

Note that the linearized dispersion relation of this equation agrees with the long wave expansion (4.48) of the linearized dispersion relation of the original system.

If a satisfies (4.54), then we may solve (4.53) for (n_2, u_2). The solution is the sum of a solution of the nonhomogeneous equations and an arbitrary multiple

$$a_2(\xi,\tau) \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

of the solution of the homogeneous problem.

We may compute higher-order terms in the asymptotic solution in a similar way. At the order ϵ^k, we obtain a nonhomogeneous linear equation for (n_k, u_k) of the form

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} n_k \\ u_k \end{pmatrix}_\xi + \begin{pmatrix} f_{k-1} \\ g_{k-1} \end{pmatrix} = 0,$$

where f_{k-1}, g_{k-1} depend only on (n_1, u_1),..., (n_{k-1}, u_{k-1}), and φ_k may be expressed explicitly in terms of $n_1, \ldots, n_k$. The condition that this equation is solvable for (n_k, u_k) is $f_{k-1} + g_{k-1} = 0$, and this condition is satisfied if a_{k-1} satisfies a suitable equation. The solution for (n_k, u_k) then involves an arbitrary function of integration a_k. An equation for a_k follows from the solvability condition for the order $(k + 1)$-equations.

In summary, the leading-order asymptotic solution of (4.47) as $\epsilon \to 0^+$ is

$$\begin{pmatrix} n \\ u \\ \varphi \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \epsilon a\left(\epsilon^{1/2}(x-t), \epsilon^{3/2} t\right) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + O(\epsilon^2),$$

where $a(\xi, \tau)$ satisfies the KdV equation (4.54). We expect that this asymptotic solution is valid for long times of the order $\tau = O(1)$ or $t = O(\epsilon^{-3/2})$.

8. Other Sturm-Liouville problems

Finally, we summarize a few other Sturm-Liouville equations and some of their applications. See [**2**] for a much more extensive list and an interesting collection of recent reviews on the subject.

8.1. Bessel's equation

This equation arises in solving the Laplace and Helmholtz equations by separation of variables in cylindrical polar coordinates:

$$-u'' + \left(\nu^2 - \frac{1}{4}\right)\frac{1}{x^2}u = \lambda u \qquad 0 < x < \infty$$

where $0 \leq \nu < \infty$ is a parameter. One pair of solutions is

$$x^{1/2}J_\nu\left(\sqrt{\lambda}x\right), \qquad x^{1/2}Y_\nu\left(\sqrt{\lambda}x\right)$$

where J_ν, Y_ν are Bessel functions of the order ν.

8.2. Legendre equations

The Legendre and associated Legendre equations arise in solving the Laplace equation in spherical polar coordinates, and give an expression for the spherical harmonic functions. The Legendre equation is

$$-\left[\left(1-x^2\right)u'\right]' + \frac{1}{4}u = \lambda u \qquad -1 < x < 1$$

The associated Legendre equation is

$$-\left[\left(1-x^2\right)u'\right]' + \frac{\mu^2}{1-x^2}u = \lambda u \qquad -1 < x < 1.$$

8.3. Laguerre equations

The Laguerre polynomials arise in solutions of the three-dimensional Schrödinger equation with an inverse-square potential, and in Gaussian integration. The Laguerre equation is

$$-\left(x^{\alpha+1}e^{-x}u'\right)' = \lambda x^\alpha e^{-x}u \qquad 0 < x < \infty,$$

where $-\infty < \alpha < \infty$.

LECTURE 5

Stochastic Processes

> We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.[1]

In many problems that involve modeling the behavior of some system, we lack sufficiently detailed information to determine how the system behaves, or the behavior of the system is so complicated that an exact description of it becomes irrelevant or impossible. In that case, a probabilistic model is often useful.

Probability and randomness have many different philosophical interpretations, but, whatever interpretation one adopts, there is a clear mathematical formulation of probability in terms of measure theory, due to Kolmogorov.

Probability is an enormous field with applications in many different areas. Here we simply aim to provide an introduction to some aspects that are useful in applied mathematics. We will do so in the context of stochastic processes of a continuous time variable, which may be thought of as a probabilistic analog of deterministic ODEs. We will focus on Brownian motion and stochastic differential equations, both because of their usefulness and the interest of the concepts they involve.

Before discussing Brownian motion in Section 3, we provide a brief review of some basic concepts from probability theory and stochastic processes.

1. Probability

> Mathematicians are like Frenchmen: whatever you say to them they translate into their own language and forthwith it is something entirely different.[2]

A *probability space* $(\Omega, \mathcal{F}, P)$ consists of: (a) a sample space Ω, whose points label all possible outcomes of a random trial; (b) a σ-algebra $\mathcal{F}$ of measurable subsets of Ω, whose elements are the events about which it is possible to obtain information; (c) a probability measure $P : \mathcal{F} \to [0, 1]$, where $0 \leq P(A) \leq 1$ is the probability that the event $A \in \mathcal{F}$ occurs. If $P(A) = 1$, we say that an event A

[1] Pierre Simon Laplace, in *A Philosophical Essay on Probabilities.*

[2] Johann Goethe. It has been suggested that Goethe should have said "Probabilists are like Frenchmen (or Frenchwomen)."

occurs *almost surely*. When the σ-algebra $\mathcal{F}$ and the probability measure P are understood from the context, we will refer to the probability space as Ω.

In this definition, we say that$\mathcal{F}$ is σ-algebra on Ω if it is is a collection of subsets of Ω such that $\emptyset$ and Ω belong to $\mathcal{F}$, the complement of a set in $\mathcal{F}$ belongs to $\mathcal{F}$, and a countable union or intersection of sets in $\mathcal{F}$ belongs to $\mathcal{F}$. A probability measure P on $\mathcal{F}$ is a function $P : \mathcal{F} \to [0,1]$ such that $P(\emptyset) = 0$, $P(\Omega) = 1$, and for any sequence $\{A_n\}$ of pairwise disjoint sets (meaning that $A_i \cap A_j = \emptyset$ for $i \neq j$) we have

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P\left(A_n\right).$$

Example 5.1. Let Ω be a set and $\mathcal{F}$ a σ-algebra on Ω. Suppose that

$$\{\omega_n \in \Omega : n \in \mathbb{N}\}$$

is a countable subset of Ω and $\{p_n\}$ is a sequence of numbers $0 \leq p_n \leq 1$ such that $p_1 + p_2 + p_3 + \cdots = 1$. Then we can define a probability measure $P : \mathcal{F} \to [0,1]$ by

$$P(A) = \sum_{\omega_n \in A} p_n.$$

If $\mathcal{E}$ is a collection of subsets of a set Ω, then the σ-algebra generated by $\mathcal{E}$, denoted $\sigma(\mathcal{E})$, is the smallest σ-algebra that contains $\mathcal{E}$.

Example 5.2. The open subsets of $\mathbb{R}$ generate a σ-algebra $\mathcal{B}$ called the Borel σ-algebra of $\mathbb{R}$. This algebra is also generated by the closed sets, or by the collection of intervals. The interval $[0,1]$ equipped with the σ-algebra $\mathcal{B}$ of its Borel subsets and Lebesgue measure, which assigns to an interval a measure equal to its length, forms a probability space. This space corresponds to the random trial of picking a uniformly distributed real number from $[0,1]$.

1.1. Random variables

A function $X : \Omega \to \mathbb{R}$ defined on a set Ω with a σ-algebra $\mathcal{F}$ is said to be $\mathcal{F}$-measurable, or simply measurable when $\mathcal{F}$ is understood, if $X^{-1}(A) \in \mathcal{F}$ for every Borel set $A \in \mathcal{B}$ in $\mathbb{R}$. A *random variable* on a probability space $(\Omega, \mathcal{F}, P)$ is a real-valued $\mathcal{F}$-measurable function $X : \Omega \to \mathbb{R}$. Intuitively, a random variable is a real-valued quantity that can be measured from the outcome of a random trial.

If $f : \mathbb{R} \to \mathbb{R}$ is a Borel measurable function, meaning that $f^{-1}(A) \in \mathcal{B}$ for every $A \in \mathcal{B}$, and X is a random variable, then $Y = f \circ X$, defined by $Y(\omega) = f\left(X(\omega)\right)$, is also a random variable.

We denote the expected value of a random variable X with respect to the probability measure P by $\mathbf{E}^P[X]$, or $\mathbf{E}[X]$ when the measure P is understood. The expected value is a real number which gives the mean value of the random variable X. Here, we assume that X is *integrable*, meaning that the expected value $\mathbf{E}[\,|X|\,] < \infty$ is finite. This is the case if large values of X occur with sufficiently low probability.

Example 5.3. If X is a random variable with mean $\mu = \mathbf{E}[X]$, the *variance* σ^2 of X is defined by

$$\sigma^2 = \mathbf{E}\left[(X-\mu)^2\right],$$

assuming it is finite. The standard deviation σ provides a measure of the departure of X from its mean μ. The *covariance* of two random variables X_1, X_2 with means

μ_1, μ_2, respectively, is defined by

$$\operatorname{cov}(X_1, X_2) = \mathbf{E}\left[(X_1 - \mu_1)(X_2 - \mu_2)\right].$$

We will also loosely refer to this quantity as a correlation function, although strictly speaking the correlation function of X_1, X_2 is equal to their covariance divided by their standard deviations.

The expectation is a linear functional on random variables, meaning that for integrable random variables X, Y and real numbers c we have

$$\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y], \qquad \mathbf{E}[cX] = c\mathbf{E}[X].$$

The expectation of an integrable random variable X may be expressed as an integral with respect to the probability measure P as

$$\mathbf{E}[X] = \int_\Omega X(\omega)\, dP(\omega).$$

In particular, the probability of an event $A \in \mathcal{F}$ is given by

$$P(A) = \int_A dP(\omega) = \mathbf{E}[1_A]$$

where $1_A : \Omega \to \{0,1\}$ is the indicator function of A,

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

We will say that two random variables are equal P-almost surely, or almost surely when P is understood, if they are equal on an event A such that $P(A) = 1$. Similarly, we say that a random variable $X : A \subset \Omega \to \mathbb{R}$ is defined almost surely if $P(A) = 1$. Functions of random variables that are equal almost surely have the same expectations, and we will usually regard such random variables as being equivalent.

Suppose that $\{X_\lambda : \lambda \in \Lambda\}$ is a collection of functions $X_\lambda : \Omega \to \mathbb{R}$. The σ-algebra generated by $\{X_\lambda : \lambda \in \Lambda\}$, denoted $\sigma(X_\lambda : \lambda \in \Lambda)$, is the smallest σ-algebra $\mathcal{G}$ such that X_λ is $\mathcal{G}$-measurable for every $\lambda \in \Lambda$. Equivalently, $\mathcal{G} = \sigma(\mathcal{E})$ where $\mathcal{E} = \left\{X_\lambda^{-1}(A) : \lambda \in \Lambda,\, A \in \mathcal{B}(\mathbb{R})\right\}$.

1.2. Absolutely continuous and singular measures

Suppose that $P, Q : \mathcal{F} \to [0,1]$ are two probability measures defined on the same σ-algebra $\mathcal{F}$ of a sample space Ω.

We say that Q is *absolutely continuous* with respect to P is there is an integrable random variable $f : \Omega \to \mathbb{R}$ such that for every $A \in \mathcal{F}$ we have

$$Q(A) = \int_A f(\omega) dP(\omega).$$

We will write this relation as

$$dQ = f dP,$$

and call f the density of Q with respect to P. It is defined P-almost surely. In that case, if $\mathbf{E}^P$ and $\mathbf{E}^Q$ denote the expectations with respect to P and Q, respectively, and X is a random variable which is integrable with respect to Q, then

$$\mathbf{E}^Q[X] = \int_\Omega X\, dQ = \int_\Omega fX\, dP = \mathbf{E}^P[fX].$$

We say that probability measures P and Q on $\mathcal{F}$ are *singular* if there is an event $A \in \mathcal{F}$ such that $P(A) = 1$ and $Q(A) = 0$ (or, equivalently, $P(A^c) = 0$ and $Q(A^c) = 1$). This means that events which occur with finite probability with respect to P almost surely do not occur with respect to Q, and visa-versa.

Example 5.4. Let P be the Lebesgue probability measure on $([0,1], \mathcal{B})$ described in Example 5.2. If $f : [0,1] \to [0,\infty)$ is a nonnegative, integrable function with

$$\int_0^1 f(\omega)\, d\omega = 1,$$

where $d\omega$ denotes integration with respect to Lebesgue measure, then we can define a measure Q on $([0,1], \mathcal{B})$ by

$$Q(A) = \int_A f(\omega)\, d\omega.$$

The measure Q is absolutely continuous with respect to P with density f. Note that P is not necessarily absolutely continuous with respect to Q; this is the case only if $f \neq 0$ almost surely and $1/f$ is integrable. If R is a measure on $([0,1], \mathcal{B})$ of the type given in Example 5.1 then R and P (or R and Q) are singular because the Lebesgue measure of any countable set is equal to zero.

1.3. Probability densities

The distribution function $F : \mathbb{R} \to [0,1]$ of a random variable $X : \Omega \to \mathbb{R}$ is defined by $F(x) = P\{\omega \in \Omega : X(\omega) \leq x\}$ or, in more concise notation,

$$F(x) = P\{X \leq x\}.$$

We say that a random variable is continuous if the probability measure it induces on $\mathbb{R}$ is absolutely continuous with respect to Lebesgue measure.[3] Most of the random variables we consider here will be continuous.

If X is a continuous random variable with distribution function F, then F is differentiable and

$$p(x) = F'(x)$$

is the probability density function of X. If $A \in \mathcal{B}(\mathbb{R})$ is a Borel subset of $\mathbb{R}$, then

$$P\{X \in A\} = \int_A p(x)\, dx.$$

The density satisfies $p(x) \geq 0$ and

$$\int_{-\infty}^{\infty} p(x)\, dx = 1.$$

Moreover, if $f : \mathbb{R} \to \mathbb{R}$ is any Borel-measurable function such that $f(X)$ is integrable, then

$$\mathbf{E}[f(X)] = \int_{-\infty}^{\infty} f(x)p(x)\, dx.$$

Example 5.5. A random variable X is Gaussian with mean μ and variance σ^2 if it has the probability density

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/\left(2\sigma^2\right)}.$$

[3]This excludes, for example, counting-type random variables that take only integer values.

We say that random variables $X_1, X_2, \ldots X_n : \Omega \to \mathbb{R}$ are jointly continuous if there is a joint probability density function $p(x_1, x_2, \ldots, x_n)$ such that

$$P\{X_1 \in A_1,\, X_1 \in A_1, \ldots,\, X_n \in A_n\} = \int_A p(x_1, x_2 \ldots, x_n)\; dx_1 dx_2 \ldots dx_n.$$

where $A = A_1 \times A_2 \times \cdots \times A_n$. Then $p(x_1, x_2, \ldots, x_n) \geq 0$ and

$$\int_{\mathbb{R}^n} p(x_1, x_2, \ldots, x_n)\; dx_1 dx_2 \ldots dx_n = 1.$$

Expected values of functions of the X_i are given by

$$\mathbf{E}[f(X_1, X_2, \ldots, X_n)] = \int_{\mathbb{R}^n} f(x_1, x_2, \ldots, x_n)\, p(x_1, x_2, \ldots, x_n)\; dx_1 dx_2 \ldots dx_n.$$

We can obtain the joint probability density of a subset of the X_i's by integrating out the other variables. For example, if $p(x, y)$ is the joint probability density of random variables X and Y, then the marginal probability densities $p_X(x)$ and $p_Y(y)$ of X and Y, respectively, are given by

$$p_X(x) = \int_{-\infty}^{\infty} p(x, y)\, dy, \qquad p_Y(y) = \int_{-\infty}^{\infty} p(x, y)\, dx.$$

Of course, in general, we cannot obtain the joint density $p(x, y)$ from the marginal densities $p_X(x)$, $p_Y(y)$, since the marginal densities do not contain any information about how X and Y are related.

Example 5.6. A random vector $\vec{X} = (X_1, \ldots, X_n)$ is Gaussian with mean $\vec{\mu} = (\mu_1, \ldots, \mu_n)$ and invertible covariance matrix $C = (C_{ij})$, where

$$\mu_i = \mathbf{E}[X_i], \qquad C_{ij} = \mathbf{E}[(X_i - \mu_i)(X_j - \mu_j)],$$

if it has the probability density

$$p(\vec{x}) = \frac{1}{(2\pi)^{n/2} (\det C)^{1/2}} \exp\left\{ -\frac{1}{2} (\vec{x} - \vec{\mu})^\top C^{-1} (\vec{x} - \vec{\mu}) \right\}.$$

Gaussian random variables are completely specified by their mean and covariance.

1.4. Independence

Random variables $X_1, X_2, \ldots, X_n : \Omega \to \mathbb{R}$ are said to be *independent* if

$$P\{X_1 \in A_1,\, X_2 \in A_2,\, \ldots,\, X_n \in A_n\} = P\{X_1 \in A_1\}\, P\{X_2 \in A_2\} \ldots P\{X_n \in A_n\}$$

for arbitrary Borel sets $A_1,\, A_2, \ldots, A_3 \subset \mathbb{R}$. If $X_1,\, X_2, \ldots,\, X_n$ are independent random variables, then

$$\mathbf{E}[f_1(X_1)\, f_2(X_2) \ldots f_n(X_n)] = \mathbf{E}[f_1(X_1)]\, \mathbf{E}[f_2(X_2)] \ldots \mathbf{E}[f_n(X_n)].$$

Jointly continuous random variables are independent if their joint probability density distribution factorizes into a product:

$$p(x_1, x_2, \ldots, x_n) = p_1(x_1)\, p_2(x_2) \ldots p_n(x_n).$$

If the densities $p_i = p_j$ are the same for every $1 \leq i, j \leq n$, then we say that $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables.

Heuristically, each random variable in a collection of independent random variables defines a different 'coordinate axis' of the probability space on which they are defined. Thus, any probability space that is rich enough to support a countably infinite collection of independent random variables is necessarily 'infinite-dimensional.'

Example 5.7. The Gaussian random variables in Example 5.6 are independent if and only if the covariance matrix C is diagonal.

The sum of independent Gaussian random variables is a Gaussian random variable whose mean and variance are the sums of those of the independent Gaussians. This is most easily seen by looking at the characteristic function of the sum,

$$\mathbf{E}\left[e^{i\xi(X_1+\cdots+X_n)}\right] = \mathbf{E}\left[e^{i\xi X_1}\right]\ldots\mathbf{E}\left[e^{i\xi X_n}\right],$$

which is the Fourier transform of the density. The characteristic function of a Gaussian with mean μ and variance σ^2 is $e^{i\xi\mu-\sigma^2\xi^2/2}$, so the means and variances add when the characteristic functions are multiplied. Also, a linear transformations of Gaussian random variables is Gaussian.

1.5. Conditional expectation

Conditional expectation is a somewhat subtle topic. We give only a brief discussion here. See [**45**] for more information and proofs of the results we state here.

First, suppose that $X : \Omega \to \mathbb{R}$ is an integrable random variable on a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra contained in $\mathcal{F}$. Then the conditional expectation of X given $\mathcal{G}$ is a $\mathcal{G}$-measurable random variable

$$\mathbf{E}\left[X \mid \mathcal{G}\right] : \Omega \to \mathbb{R}$$

such that for all bounded $\mathcal{G}$-measurable random variables Z

$$\mathbf{E}\left[\,\mathbf{E}\left[\,X \mid \mathcal{G}\right] Z\,\right] = \mathbf{E}\left[XZ\right].$$

In particular, choosing $Z = 1_B$ as the indicator function of $B \in \mathcal{G}$, we get

$$\int_B \mathbf{E}\left[X \mid \mathcal{G}\right] dP = \int_B X\, dP \qquad \text{for all } B \in \mathcal{G}. \tag{5.1}$$

The existence of $\mathbf{E}\left[X \mid \mathcal{G}\right]$ follows from the Radon-Nikodym theorem or by a projection argument. The conditional expectation is only defined up to almost-sure equivalence, since (5.1) continues to hold if $\mathbf{E}[X \mid \mathcal{G}]$ is modified on an event in $\mathcal{G}$ that has probability zero. Any equations that involve conditional expectations are therefore understood to hold almost surely.

Equation (5.1) states, roughly, that $\mathbf{E}\left[X \mid \mathcal{G}\right]$ is obtained by averaging X over the finer σ-algebra $\mathcal{F}$ to get a function that is measurable with respect to the coarser σ-algebra $\mathcal{G}$. Thus, one may think of $\mathbf{E}\left[X \mid \mathcal{G}\right]$ as providing the 'best' estimate of X given information about the events in $\mathcal{G}$.

It follows from the definition that if X, XY are integrable and Y is $\mathcal{G}$-measurable then

$$\mathbf{E}\left[XY \mid \mathcal{G}\right] = Y\mathbf{E}\left[X \mid \mathcal{G}\right].$$

Example 5.8. The conditional expectation given the full σ-algebra $\mathcal{F}$, corresponding to complete information about events, is $\mathbf{E}\left[X \mid \mathcal{F}\right] = X$. The conditional expectation given the trivial σ-algebra $\mathcal{M} = \{\emptyset, \Omega\}$, corresponding to no information about events, is the constant function $\mathbf{E}\left[X \mid \mathcal{G}\right] = \mathbf{E}[X]$.

Example 5.9. Suppose that $\mathcal{G} = \{\emptyset, B, B^c, \Omega\}$ where B is an event such that $0 < P(B) < 1$. This σ-algebra corresponds to having information about whether or not the event B has occurred. Then

$$\mathbf{E}\left[X \mid \mathcal{G}\right] = p1_B + q1_{B^c}$$

where p, q are the expected values of X on B, B^c, respectively

$$p = \frac{1}{P(B)} \int_B X \, dP, \qquad q = \frac{1}{P(B^c)} \int_{B^c} X \, dP.$$

Thus, $\mathbf{E}[X \mid \mathcal{G}](\omega)$ is equal to the expected value of X given B if $\omega \in B$, and the expected value of X given B^c if $\omega \in B^c$.

The conditional expectation has the following 'tower' property regarding the collapse of double expectations into a single expectation: If $\mathcal{H} \subset \mathcal{G}$ are σ-algebras, then

$$\mathbf{E}[\mathbf{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathbf{E}[\mathbf{E}[X \mid \mathcal{H}] \mid \mathcal{G}] = \mathbf{E}[X \mid \mathcal{H}], \tag{5.2}$$

sometimes expressed as 'the coarser algebra wins.'

If $X, Y : \Omega \to \mathbb{R}$ are integrable random variables, we define the conditional expectation of X given Y by

$$\mathbf{E}[X \mid Y] = \mathbf{E}[X \mid \sigma(Y)].$$

This random variable depends only on the events that Y defines, not on the values of Y themselves.

Example 5.10. Suppose that $Y : \Omega \to \mathbb{R}$ is a random variable that attains countably many distinct values y_n. The sets $B_n = Y^{-1}(y_n)$, form a countable disjoint partition of Ω. For any integrable random variable X, we have

$$\mathbf{E}[X \mid Y] = \sum_{n \in \mathbb{N}} z_n \, 1_{B_n}$$

where 1_{B_n} is the indicator function of B_n, and

$$z_n = \frac{\mathbf{E}[1_{B_n} X]}{P(B_n)} = \frac{1}{P(B_n)} \int_{B_n} X \, dP$$

is the expected value of X on B_n. Here, we assume that $P(B_n) \neq 0$ for every $n \in \mathbb{N}$. If $P(B_n) = 0$ for some n, then we omit that term from the sum, which amounts to defining $\mathbf{E}[X \mid Y](\omega) = 0$ for $\omega \in B_n$. The choice of a value other than 0 for $\mathbf{E}[X \mid Y]$ on B_n would give an equivalent version of the conditional expectation. Thus, if $Y(\omega) = y_n$ then $\mathbf{E}[X \mid Y](\omega) = z_n$ where z_n is the expected value of $X(\omega')$ over all ω' such that $Y(\omega') = y_n$. This expression for the conditional expectation does not apply to continuous random variables Y, since then $P\{Y = y\} = 0$ for every $y \in \mathbb{R}$, but we will give analogous results below for continuous random variables in terms of their probability densities.

If $Y, Z : \Omega \to \mathbb{R}$ are random variables such that Z is measurable with respect to $\sigma(Y)$, then one can show that there is a Borel function $\varphi : \mathbb{R} \to \mathbb{R}$ such that $Z = \varphi(Y)$. Thus, there is a Borel function $\varphi : \mathbb{R} \to \mathbb{R}$ such that

$$\mathbf{E}[X \mid Y] = \varphi(Y).$$

We then define the conditional expectation of X given that $Y = y$ by

$$\mathbf{E}[X \mid Y = y] = \varphi(y).$$

Since the conditional expectation $\mathbf{E}[X \mid Y]$ is, in general, defined almost surely, we cannot define $\mathbf{E}[X \mid Y = y]$ unambiguously for all $y \in \mathbb{R}$, only for $y \in A$ where A is a Borel subset of $\mathbb{R}$ such that $P\{Y \in A\} = 1$.

More generally, if $Y_1, \ldots, Y_n$ are random variables, we define the conditional expectation of an integrable random variable X given $Y_1, \ldots, Y_n$ by

$$\mathbf{E}\left[X \mid Y_1, \ldots, Y_n\right] = \mathbf{E}\left[X \mid \sigma\left(Y_1, \ldots, Y_n\right)\right].$$

This is a random variable $\mathbf{E}\left[X \mid Y_1, \ldots, Y_n\right] : \Omega \to \mathbb{R}$ which is measurable with respect to $\sigma\left(Y_1, \ldots, Y_n\right)$ and defined almost surely. As before, there is a Borel function $\varphi : \mathbb{R}^n \to \mathbb{R}$ such that $\mathbf{E}\left[X \mid Y_1, \ldots, Y_n\right] = \varphi\left(Y_1, \ldots, Y_n\right)$. We denote the corresponding conditional expectation of X given that $Y_1 = y_1, \ldots, Y_n = y_n$ by

$$\mathbf{E}\left[X \mid Y_1 = y_1, \ldots, Y_n = y_n\right] = \varphi\left(y_1, \ldots, y_n\right).$$

Next we specialize these results to the case of continuous random variables. Suppose that $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are random variables with a joint probability density $p\left(x_1, x_2, \ldots x_m, y_1, y_2, \ldots, y_n\right)$. The conditional joint probability density of X_1, X_2,..., X_m given that $Y_1 = y_1$, $Y_2 = y_2$,..., $Y_n = y_n$, is

$$p\left(x_1, x_2, \ldots x_m \mid y_1, y_2, \ldots, y_n\right) = \frac{p\left(x_1, x_2, \ldots x_m, y_1, y_2, \ldots, y_n\right)}{p_Y\left(y_1, y_2, \ldots, y_n\right)}, \tag{5.3}$$

where p_Y is the marginal density of the $(Y_1, \ldots, Y_n)$,

$$p_Y\left(y_1, \ldots, y_n\right) = \int_{\mathbb{R}^m} p\left(x_1, \ldots, x_m, y_1, \ldots, y_n\right)\, dx_1 \ldots dx_m.$$

The conditional expectation of $f\left(X_1, \ldots, X_m\right)$ given that $Y_1 = y_1, \ldots, Y_n = y_n$ is

$$\begin{aligned}&\mathbf{E}\left[f\left(X_1, \ldots, X_m\right) \mid Y_1 = y_1, \ldots, Y_n = y_n\right]\\ &\quad = \int_{\mathbb{R}^m} f\left(x_1, \ldots, x_m\right) p\left(x_1, \ldots x_m \mid y_1, \ldots, y_n\right)\, dx_1, \ldots, dx_m.\end{aligned}$$

The conditional probability density $p\left(x_1, \ldots x_m \mid y_1, \ldots, y_n\right)$ in (5.3) is defined for $(y_1, \ldots, y_n) \in A$, where $A = \{(y_1, \ldots, y_n) \in \mathbb{R}^n : p_Y\left(y_1, \ldots, y_n\right) > 0\}$. Since

$$P\left\{(Y_1, \ldots, Y_n) \in A^c\right\} = \int_{A^c} p_Y(y_1, \ldots, y_n)\, dy_1 \ldots dy_n = 0$$

we have $P\{(Y_1, \ldots, Y_n) \in A\} = 1$.

Example 5.11. If X, Y are random variables with joint probability density $p(x, y)$, then the conditional probability density of X given that $Y = y$, is defined by

$$p(x \mid y) = \frac{p(x, y)}{p_Y(y)}, \qquad p_Y(y) = \int_{-\infty}^{\infty} p(x, y)\, dx,$$

provided that $p_Y(y) > 0$. Also,

$$\mathbf{E}\left[f(X, Y) \mid Y = y\right] = \int_{-\infty}^{\infty} f(x, y) p(x \mid y)\, dx = \frac{\int_{-\infty}^{\infty} f(x, y) p(x, y)\, dx}{p_Y(y)}.$$

2. Stochastic processes

Consider a real-valued quantity that varies 'randomly' in time. For example, it could be the brightness of a twinkling star, a velocity component of the wind at a weather station, a position or velocity coordinate of a pollen grain in Brownian motion, the number of clicks recorded by a Geiger counter up to a given time, or the value of the Dow-Jones index.

We describe such a quantity by a measurable function

$$X : [0, \infty) \times \Omega \to \mathbb{R}$$

where Ω is a probability space, and call X a stochastic process. The quantity $X(t,\omega)$ is the value of the process at time t for the outcome $\omega \in \Omega$. When it is not necessary to refer explicitly to the dependence of $X(t,\omega)$ on ω, we will write the process as $X(t)$. We consider processes that are defined on $0 \le t < \infty$ for definiteness, but one can also consider processes defined on other time intervals, such as $[0,1]$ or $\mathbb{R}$. One can also consider discrete-time processes with $t \in \mathbb{N}$, or $t \in \mathbb{Z}$, for example. We will consider only continuous-time processes.

We may think of a stochastic process in two different ways. First, fixing $\omega \in \Omega$, we get a function of time

$$X^\omega : t \mapsto X(t,\omega),$$

called a sample function (or sample path, or realization) of the process. From this perspective, the process is a collection of functions of time $\{X^\omega : \omega \in \Omega\}$, and the probability measure is a measure on the space of sample functions.

Alternatively, fixing $t \in [0,\infty)$, we get a random variable

$$X_t : \omega \mapsto X(t,\omega)$$

defined on the probability space Ω. From this perspective, the process is a collection of random variables $\{X_t : 0 \le t < \infty\}$ indexed by the time variable t. The probability measure describes the joint distribution of these random variables.

2.1. Distribution functions

A basic piece of information about a stochastic process X is the probability distribution of the random variables X_t for each $t \in [0,\infty)$. For example if X_t is continuous, we can describe its distribution by a probability density $p(x,t)$. These one-point distributions do not, however, tell us how the values of the process at different times are related.

Example 5.12. Let X be a process such that with probability $1/2$, we have $X_t = 1$ for all t, and with probability $1/2$, we have $X_t = -1$ for all t. Let Y be a process such that Y_t and Y_s are independent random variables for $t \neq s$, and for each t, we have $Y_t = 1$ with probability $1/2$ and $Y_t = -1$ with probability $1/2$. Then X_t, Y_t have the same distribution for each $t \in \mathbb{R}$, but they are different processes, because the values of X at different times are completely correlated, while the values of Y are independent. As a result, the sample paths of X are constant functions, while the sample paths of Y are almost surely discontinuous at every point (and non-Lebesgue measurable). The means of these processes, $\mathbf{E}X_t = \mathbf{E}Y_t = 0$, are equal and constant, but they have different covariances

$$\mathbf{E}\left[X_s X_t\right] = 1, \qquad \mathbf{E}\left[Y_s Y_t\right] = \begin{cases} 1 & \text{if } t = s, \\ 0 & \text{otherwise.} \end{cases}$$

To describe the relationship between the values of a process at different times, we need to introduce multi-dimensional distribution functions. We will assume that the random variables associated with the process are continuous.

Let $0 \le t_1 < t_2 < \cdots < t_n$ be a sequence times, and A_1, A_2,... A_n a sequence of Borel subsets $\mathbb{R}$. Let E be the event

$$E = \left\{\omega \in \Omega : X_{t_j}(\omega) \in A_j \text{ for } 1 \le j \le n\right\}.$$

Then, assuming the existence of a joint probability density $p\,(x_n, t_;\ldots;x_2, t_2; x_1, t_1)$ for X_{t_1}, X_{t_2},..., X_{t_n}, we can write

$$P\{E\} = \int_A p\,(x_n, t_n;\ldots;x_2, t_2; x_1, t_1)\; dx_1 dx_2 \ldots dx_n$$

where $A = A_1 \times A_2 \times \cdots \times A_n \subset \mathbb{R}^n$. We adopt the convention that times are written in increasing order from right to left in p.

These finite-dimensional densities must satisfy a consistency condition relating the $(n+1)$-dimensional densities to the n-dimensional densities: If $n \in \mathbb{N}$, $1 \leq i \leq n$ and $t_1 < t_2 < \cdots < t_i < \cdots < t_n$, then

$$\int_{-\infty}^{\infty} p\,(x_{n+1}, t_{n+1};\ldots;x_{i+1}, t_{i+1}; x_i, t_i; x_{i-1}, t_{i-1};\ldots;x_1, t_1)\; dx_i$$
$$= p\,(x_{n+1}, t_{n+1};\ldots;x_{i+1}, t_{i+1}; x_{i-1}, t_{i-1};\ldots;x_1, t_1)\,.$$

We will regard these finite-dimensional probability densities as providing a full description of the process. For continuous-time processes this requires an assumption of separability, meaning that the process is determined by its values at countably many times. This is the case, for example, if its sample paths are continuous, so that they are determined by their values at all rational times.

Example 5.13. To illustrate the inadequacy of finite-dimensional distributions for the description of non-separable processes, consider the process $X : [0,1] \times \Omega \to \mathbb{R}$ defined by

$$X(t,\omega) = \begin{cases} 1 & \text{if } t = \omega, \\ 0 & \text{otherwise,} \end{cases}$$

where $\Omega = [0,1]$ and P is Lebesgue measure on Ω. In other words, we pick a point $\omega \in [0,1]$ at random with respect to a uniform distribution, and change X_t from zero to one at $t = \omega$. The single time distribution of X_t is given by

$$P\,\{X_t \in A\} = \begin{cases} 1 & \text{if } 0 \in A, \\ 0 & \text{otherwise,} \end{cases}$$

since the probability that $\omega = t$ is zero. Similarly,

$$P\,\{X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n\} = \begin{cases} 1 & \text{if } 0 \in \bigcap_{i=1}^n A_i, \\ 0 & \text{otherwise,} \end{cases}$$

since the probability that $\omega = t_i$ for some $1 \leq i \leq n$ is also zero. Thus, X has the same finite-dimensional distributions as the trivial zero-process $Z(t,\omega) = 0$. If, however, we ask for the probability that the realizations are continuous, we get different answers:

$$P\,\{X^\omega \text{ is continuous on } [0,1]\} = 0, \qquad P\,\{Z^\omega \text{ is continuous on } [0,1]\} = 1.$$

The problem here is that in order to detect the discontinuity in a realization X^ω of X, one needs to look at its values at an uncountably infinite number of times. Since measures are only countably additive, we cannot determine the probability of such an event from the probability of events that depend on the values of X^ω at a finite or countably infinite number of times.

2.2. Stationary processes

A process X_t, defined on $-\infty < t < \infty$, is *stationary* if X_{t+c} has the same distribution as X_t for all $-\infty < c < \infty$; equivalently this means that all of its finite-dimensional distributions depend only on time differences. 'Stationary' here is used in a probabilistic sense; it does not, of course, imply that the individual sample functions do not vary in time. For example, if one considers the fluctuations of a thermodynamic quantity, such as the pressure exerted by a gas on the walls of its container, this quantity varies in time even when the system is in thermodynamic equilibrium. The one-point probability distribution of the quantity is independent of time, but the two-point correlation at different times depends on the time difference.

2.3. Gaussian processes

A process is *Gaussian* if all of its finite-dimensional distributions are multivariate Gaussian distributions. A separable Gaussian process is completely determined by the means and covariance matrices of its finite-dimensional distributions.

2.4. Filtrations

Suppose that $X : [0, \infty) \times \Omega \to \mathbb{R}$ is a stochastic process on a probability space Ω with σ-algebra $\mathcal{F}$. For each $0 \leq t < \infty$, we define a σ-algebra $\mathcal{F}_t$ by

$$\mathcal{F}_t = \sigma\left(X_s : 0 \leq s \leq t\right). \tag{5.4}$$

If $0 \leq s < t$, then $\mathcal{F}_s \subset \mathcal{F}_t \subset \mathcal{F}$. Such a family of σ-fields $\{\mathcal{F}_t : 0 \leq t < \infty\}$ is called a *filtration* of $\mathcal{F}$.

Intuitively, $\mathcal{F}_t$ is the collection of events whose occurrence can be determined from observations of the process up to time t, and an $\mathcal{F}_t$-measurable random variable is one whose value can be determined by time t. If X is any random variable, then $\mathbf{E}\left[\, X \mid \mathcal{F}_t \,\right]$ is the 'best' estimate of X based on observations of the process up to time t.

The properties of conditional expectations with respect to filtrations define various types of stochastic processes, the most important of which for us will be Markov processes.

2.5. Markov processes

A stochastic process X is said to be a *Markov process* if for any $0 \leq s < t$ and any Borel measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $f(X_t)$ has finite expectation, we have

$$\mathbf{E}\left[f\left(X_t\right) \mid \mathcal{F}_s\right] = \mathbf{E}\left[f\left(X_t\right) \mid X_s\right].$$

Here $\mathcal{F}_s$ is defined as in (5.4). This property means, roughly, that 'the future is independent of the past given the present.' In anthropomorphic terms, a Markov process only cares about its present state, and has no memory of how it got there.

We may also define a Markov process in terms of its finite-dimensional distributions. As before, we consider only processes for which the random variables X_t are continuous, meaning that their distributions can be described by probability densities. For any times

$$0 \leq t_1 < t_2 < \cdots < t_m < t_{m+1} < \cdots < t_n,$$

the conditional probability density that $X_{t_i} = x_i$ for $m+1 \le i \le n$ given that $X_{t_i} = x_i$ for $1 \le i \le m$ is given by

$$p\left(x_n, t_n; \ldots; x_{m+1}, t_{m+1} \mid x_m, t_m; \ldots; x_1, t_1\right) = \frac{p\left(x_n, t_n; \ldots; x_1, t_1\right)}{p\left(x_m, t_m; \ldots; x_1, t_1\right)}.$$

The process is a Markov process if these conditional densities depend only on the conditioning at the most recent time, meaning that

$$p\left(x_{n+1}, t_{n+1} \mid x_n, t_n; \ldots; x_2, t_2; x_1, t_1\right) = p\left(x_{n+1}, t_{n+1} \mid x_n, t_n\right).$$

It follows that, for a Markov process,

$$p\left(x_n, t_n; \ldots; x_2, t_2 \mid x_1, t_1\right) = p\left(x_n, t_n \mid x_{n-1}, t_{n-1}\right) \ldots p\left(x_2, t_2 \mid x_1, t_1\right).$$

Thus, we can determine all joint finite-dimensional probability densities of a continuous Markov process X_t in terms of the transition density $p\left(x, t \mid y, s\right)$ and the probability density $p_0(x)$ of its initial value X_0. For example, the one-point density of X_t is given by

$$p(x,t) = \int_{-\infty}^{\infty} p\left(x, t \mid y, 0\right) p_0(y)\, dy.$$

The transition probabilities of a Markov process are not arbitrary and satisfy the *Chapman-Kolmogorov equation*. In the case of a continuous Markov process, this equation is

$$p(x,t \mid y,s) = \int_{-\infty}^{\infty} p(x,t \mid z,r) p(z,r \mid y,s)\, dz \qquad \text{for any } s < r < t, \tag{5.5}$$

meaning that in going from y at time s to x at time t, the process must go though some point z at any intermediate time r.

A continuous Markov process is *time-homogeneous* if

$$p(x,t \mid y,s) = p(x,t-s \mid y,0),$$

meaning that its stochastic properties are invariant under translations in time. For example, a stochastic differential equation whose coefficients do not depend explicitly on time defines a time-homogeneous continuous Markov process. In that case, we write $p(x,t \mid y,s) = p(x,t-s \mid y)$ and the Chapman-Kolmogorov equation (5.5) becomes

$$p(x,t \mid y) = \int_{-\infty}^{\infty} p(x,t-s \mid z) p(z,s \mid y)\, dz \qquad \text{for any } 0 < s < t. \tag{5.6}$$

Nearly all of the processes we consider will be time-homogeneous.

2.6. Martingales

Martingales are fundamental to the analysis of stochastic processes, and they have important connections with Brownian motion and stochastic differential equations. Although we will not make use of them, we give their definition here.

We restrict our attention to processes M with continuous sample paths on a probability space $(\Omega, \mathcal{F}_t, P)$, where $\mathcal{F}_t = \sigma\left(M_t : t \ge 0\right)$ is the filtration induced by M. Then M is a *martingale*[4] if M_t has finite expectation for every $t \ge 0$ and for

[4]The term 'martingale' was apparently used in 18th century France as a name for the roulette betting 'strategy' of doubling the bet after every loss. If one were to compile a list of nondescriptive and off-putting names for mathematical concepts, 'martingale' would almost surely be near the top.

any $0 \leq s < t$,

$$\mathbf{E}[\, M_t \mid \mathcal{F}_s \,] = M_s.$$

Intuitively, a martingale describes a 'fair game' in which the expected value of a player's future winnings M_t is equal to the player's current winnings M_s. For more about martingales, see [**46**], for example.

3. Brownian motion

> The grains of pollen were particles...of a figure between cylindrical and oblong, perhaps slightly flattened...While examining the form of these particles immersed in water, I observed many of them very evidently in motion; their motion consisting not only of a change in place in the fluid manifested by alterations in their relative positions...In a few instances the particle was seen to turn on its longer axis. These motions were such as to satisfy me, after frequently repeated observations, that they arose neither from currents in the fluid, nor from its gradual evaporation, but belonged to the particle itself.[5]

In 1827, Robert Brown observed that tiny pollen grains in a fluid undergo a continuous, irregular movement that never stops. Although Brown was perhaps not the first person to notice this phenomenon, he was the first to study it carefully, and it is now known as Brownian motion.

The constant irregular movement was explained by Einstein (1905) and the Polish physicist Smoluchowski (1906) as the result of fluctuations caused by the bombardment of the pollen grains by liquid molecules. (It is not clear that Einstein was initially aware of Brown's observations — his motivation was to look for phenomena that could provide evidence of the atomic nature of matter.)

For example, a colloidal particle of radius $10^{-6}\,\mathrm{m}$ in a liquid, is subject to approximately 10^{20} molecular collisions each second, each of which changes its velocity by an amount on the order of $10^{-8}\,\mathrm{m\,s^{-1}}$. The effect of such a change is imperceptible, but the cumulative effect of an enormous number of impacts leads to observable fluctuations in the position and velocity of the particle.[6]

Einstein and Smoluchowski adopted different approaches to modeling this problem, although their conclusions were similar. Einstein used a general, probabilistic argument to derive a diffusion equation for the number density of Brownian particles as a function of position and time, while Smoluchowski employed a detailed kinetic model for the collision of spheres, representing the molecules and the Brownian particles. These approaches were partially connected by Langevin (1908) who introduced the Langevin equation, described in Section 5 below.

Perrin (1908) carried out experimental observations of Brownian motion and used the results, together with Einstein's theoretical predictions, to estimate Avogadro's number N_A; he found $N_A \approx 7 \times 10^{23}$ (see Section 6.2). Thus, Brownian motion provides an almost direct observation of the atomic nature of matter.

Independently, Louis Bachelier (1900), in his doctoral dissertation, introduced Brownian motion as a model for asset prices in the French bond market. This work received little attention at the time, but there has been extensive subsequent use of

[5]Robert Brown, from *Miscellaneous Botanical Works* Vol. I, 1866.

[6]Deutsch (1992) suggested that these fluctuations are in fact too small for Brown to have observed them with contemporary microscopes, and that the motion Brown saw had some other cause.

the theory of stochastic processes to model financial markets, especially following the development of the Black-Scholes-Merton (1973) model for options pricing (see Section 8).

Wiener (1923) gave the first construction of Brownian motion as a measure on the space of continuous functions, now called Wiener measure. Wiener did this by several different methods, including the use of Fourier series with random coefficients (*c.f.* (5.7) below). This work was further developed by Wiener and many others, especially Lévy (1939).

3.1. Definition

Standard (one-dimensional) *Brownian motion* starting at 0, also called the *Wiener process*, is a stochastic process $B(t, \omega)$ with the following properties:

(1) $B(0, \omega) = 0$ for every $\omega \in \Omega$;
(2) for every $0 \leq t_1 < t_2 < t_3 < \cdots < t_n$, the increments
$$B_{t_2} - B_{t_1}, \quad B_{t_3} - B_{t_2}, \ldots, \quad B_{t_n} - B_{t_{n-1}}$$
are independent random variables;
(3) for each $0 \leq s < t < \infty$, the increment $B_t - B_s$ is a Gaussian random variable with mean 0 and variance $t - s$;
(4) the sample paths $B^\omega : [0, \infty) \to \mathbb{R}$ are continuous functions for every $\omega \in \Omega$.

The existence of Brownian motion is a non-trivial fact. The main issue is to show that the Gaussian probability distributions, which imply that $B(t+\Delta t) - B(t)$ is typically of the order $\sqrt{\Delta t}$, are consistent with the continuity of sample paths. We will not give a proof here, or derive the properties of Brownian motion, but we will describe some results which give an idea of how it behaves. For more information on the rich mathematical theory of Brownian motion, see for example [**15, 46**].

The Gaussian assumption must, in fact, be satisfied by any process with independent increments and continuous sample sample paths. This is a consequence of the central limit theorem, because each increment
$$B_t - B_s = \sum_{i=0}^{n} \left(B_{t_{i+1}} - B_{t_i} \right) \qquad s = t_0 < t_1 < \cdots < t_n = t,$$
is a sum of arbitrarily many independent random variables with zero mean; the continuity of sample paths is sufficient to ensure that the hypotheses of the central limit theorem are satisfied. Moreover, since the means and variances of independent Gaussian variables are additive, they must be linear functions of the time difference. After normalization, we may assume that the mean of $B_t - B_s$ is zero and the variance is $t - s$, as in standard Brownian motion.

Remark 5.14. A probability distribution F is said to be *infinitely divisible* if, for every $n \in \mathbb{N}$, there exists a probability distribution F_n such that if $X_1, \ldots X_n$ are independent, identically distributed random variables with distribution F_n, then $X_1 + \cdots + X_n$ has distribution F. The Gaussian distribution is infinitely divisible, since a Gaussian random variable with mean μ and variance σ^2 is a sum of n independent, identically distributed random variables with mean μ/n and variance σ^2/n, but it is not the only such distribution; the Poisson distribution is another basic example. One can construct a stochastic process with independent increments for any infinitely divisible probability distribution. These processes are called Lévy

processes [**5**]. Brownian motion is, however, the only Lévy process whose sample paths are almost surely continuous; the paths of other Lévy processes contain jump discontinuities in any time interval with nonzero probability.

Since Brownian motion is a sum of arbitrarily many independent increments in any time-interval, it has a random fractal structure in which any part of the motion, after rescaling, has the same distribution as the original motion (see Figure 1). Specifically, if $c > 0$ is a constant, then

$$\tilde{B}_t = \frac{1}{c^{1/2}} B_{ct}$$

has the same distribution as B_t, so it is also a Brownian motion. Moreover, we may translate a Brownian motion B_t from any time s back to the origin to get a Brownian motion $\hat{B}_t = B_{t+s} - B_s$, and then rescale the translated process.

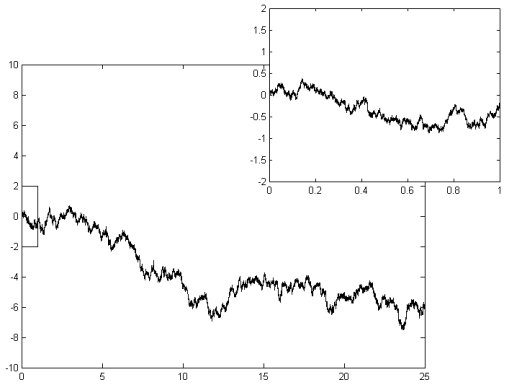

FIGURE 1. A sample path for Brownian motion, and a rescaling of it near the origin to illustrate the random fractal nature of the paths.

The condition of independent increments implies that Brownian motion is a Gaussian Markov process. It is not, however, stationary; for example, the variance t of B_t is not constant and grows linearly in time. We will discuss a closely related process in Section 5, called the stationary Ornstein-Uhlenbeck process, which is a stationary, Gaussian, Markov process (in fact, it is the only such process in one space dimension with continuous sample paths).

One way to think about Brownian motion is as a limit of random walks in discrete time. This provides an analytical construction of Brownian motion, and can be used to simulate it numerically. For example, consider a particle on a line that starts at $x = 0$ when $t = 0$ and moves as follows: After each time interval of length Δt, it steps a random distance sampled from independent, identically distributed Gaussian distributions with mean zero and variance Δt. Then, according to Donsker's theorem, the random walk approaches a Brownian motion in distribution as $\Delta t \to 0$. A key point is that although the total distance moved by the particle after time t goes to infinity as $\Delta t \to 0$, since it takes roughly on the order of $1/\Delta t$ steps of size $\sqrt{\Delta t}$, the net distance traveled remains finite almost surely

because of the cancelation between forward and backward steps, which have mean zero.

Another way to think about Brownian motion is in terms of random Fourier series. For example, Wiener (1923) showed that if $A_0, A_1, \ldots, A_n, \ldots$ are independent, identically distributed Gaussian variables with mean zero and variance one, then the Fourier series

$$B(t) = \frac{1}{\sqrt{\pi}} \left(A_0 t + 2 \sum_{n=1}^{\infty} A_n \frac{\sin nt}{n} \right) \tag{5.7}$$

almost surely has a subsequence of partial sums that converges uniformly to a continuous function. Furthermore, the resulting process B is a Brownian motion on $[0, \pi]$. The n^{th} Fourier coefficient in (5.7) is typically of the order $1/n$, so the uniform convergence of the series depends essentially on the cancelation between terms that results from the independence of their random coefficients.

3.2. Probability densities and the diffusion equation

Next, we consider the description of Brownian motion in terms of its finite-dimensional probability densities. Brownian motion is a time-homogeneous Markov process, with transition density

$$p(x, t \mid y) = \frac{1}{\sqrt{2\pi t}} \, e^{-(x-y)^2/2t} \qquad \text{for } t > 0. \tag{5.8}$$

As a function of (x, t), the transition density satisfies the diffusion, or heat, equation

$$\frac{\partial p}{\partial t} = \frac{1}{2} \frac{\partial^2 p}{\partial x^2}, \tag{5.9}$$

and the initial condition

$$p(x, 0 \mid y) = \delta(x - y).$$

The one-point probability density for Brownian motion starting at 0 is the Green's function of the diffusion equation,

$$p(x, t) = \frac{1}{\sqrt{2\pi t}} \, e^{-x^2/2t}.$$

More generally, if a Brownian motion B_t does not start almost surely at 0 and the initial value B_0 is a continuous random variable, independent of the rest of the motion, with density $p_0(x)$, then the density of B_t for $t > 0$ is given by

$$p(x, t) = \frac{1}{\sqrt{2\pi t}} \int e^{-(x-y)^2/2t} p_0(y) \, dy. \tag{5.10}$$

This is the Green's function representation of the solution of the diffusion equation (5.9) with initial data $p(x, 0) = p_0(x)$.

One may verify explicitly that the transition density (5.8) satisfies the Chapman-Kolmogorov equation (5.6). If we introduce the solution operators of (5.9),

$$T_t : p_0(\cdot) \mapsto p(\cdot, t)$$

defined by (5.10), then the Chapman-Kolmogorov equation is equivalent to the semi-group property $T_t T_s = T_{t+s}$. We use the term 'semi-group' here, because we cannot, in general, solve the diffusion equation backward in time, so T_t does not have an inverse (as would be required in a group).

The covariance function of Brownian motion is given by

$$\mathbf{E}\left[B_t B_s\right] = \min(t, s). \tag{5.11}$$

To see this, suppose that $s < t$. Then the increment $B_t - B_s$ has zero mean and is independent of B_s, and B_s has variance s, so

$$\mathbf{E}\left[B_s B_t\right] = \mathbf{E}\left[\left(B_t - B_s\right) B_s\right] + \mathbf{E}\left[B_s^2\right] = s.$$

Equivalently, we may write (5.11) as

$$\mathbf{E}\left[B_s B_t\right] = \frac{1}{2}\left(|t| + |s| - |t-s|\right).$$

Remark 5.15. One can a define a Gaussian process X_t, depending on a parameter $0 < H < 1$, called *fractional Brownian motion* which has mean zero and covariance function

$$\mathbf{E}\left[X_s X_t\right] = \frac{1}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right).$$

The parameter H is called the Hurst index of the process. When $H = 1/2$, we get Brownian motion. This process has similar fractal properties to standard Brownian motion because of the scaling-invariance of its covariance [**17**].

3.3. Sample path properties

Although the sample paths of Brownian motion are continuous, they are almost surely non-differentiable at every point.

We can describe the non-differentiablity of Brownian paths more precisely. A function $F : [a,b] \to \mathbb{R}$ is Hölder continuous on the interval $[a,b]$ with exponent γ, where $0 < \gamma \leq 1$, if there exists a constant C such that

$$|F(t) - F(s)| \leq C|t-s|^{\gamma} \qquad \text{for all } s,t \in [a,b].$$

For $0 < \gamma < 1/2$, the sample functions of Brownian motion are almost surely Hölder continuous with exponent γ on every bounded interval; but for $1/2 \leq \gamma \leq 1$, they are almost surely not Hölder continuous with exponent γ on any bounded interval.

One way to understand these results is through the law of the iterated logarithm, which states that, almost surely,

$$\limsup_{t\to 0^+} \frac{B_t}{\left(2t\log\log\frac{1}{t}\right)^{1/2}} = 1, \qquad \liminf_{t\to 0^+} \frac{B_t}{\left(2t\log\log\frac{1}{t}\right)^{1/2}} = -1.$$

Thus, although the typical fluctuations of Brownian motion over times Δt are of the order $\sqrt{\Delta t}$, there are rare deviations which are larger by a very slowly growing, but unbounded, double-logarithmic factor of $\sqrt{2\log\log(1/\Delta t)}$.

Although the sample paths of Brownian motion are almost surely not Hölder continuous with exponent $1/2$, there is a sense in which Brownian motion satisfies a stronger condition probabilistically: When measured with respect to a given, non-random set of partitions, the quadratic variation of a Brownian path on an interval of length t is almost surely equal to t. This property is of particular significance in connection with Itô's theory of stochastic differential equations (SDEs).

In more detail, suppose that $[a,b]$ is any time interval, and let $\{\Pi_n : n \in \mathbb{N}\}$ be a sequence of non-random partitions of $[a,b]$,

$$\Pi_n = \{t_0, t_1, \ldots, t_n\}, \qquad a = t_0 < t_1 < \cdots < t_n = b.$$

To be specific, suppose that Π_n is obtained by dividing $[a,b]$ into n subintervals of equal length (the result is independent of the choice of the partitions, provided they are not allowed to depend on $\omega \in \Omega$ so they cannot be 'tailored' to fit each

realization individually). We define the *quadratic variation* of a sample function B_t on the time-interval $[a, b]$ by

$$QV_a^b(B_t) = \lim_{n\to\infty} \sum_{i=1}^{n} \left(B_{t_i} - B_{t_{i-1}}\right)^2.$$

The n terms in this sum are independent, identically distributed random variables with mean $(b-a)/n$ and variance $2(b-a)^2/n^2$. Thus, the sum has mean $(b-a)$ and variance proportional to $1/n$. Therefore, by the law of large numbers, the limit exists almost surely and is equal to $(b-a)$. By contrast, the quadratic variation of any continuously differentiable function, or any function of bounded variation, is equal to zero.

This property of Brownian motion leads to the formal rule of the Itô calculus that

$$(dB)^2 = dt. \tag{5.12}$$

The apparent peculiarity of this formula, that the 'square of an infinitesimal' is another first-order infinitesimal, is a result of the nonzero quadratic variation of the Brownian paths.

The Hölder continuity of the Brownian sample functions for $0 < \gamma < 1/2$ implies that, for any $\alpha > 2$, the α-variation is almost surely equal to zero:

$$\lim_{n\to\infty} \sum_{i=1}^{n} \left|B_{t_i} - B_{t_{i-1}}\right|^{\alpha} = 0.$$

3.4. Wiener measure

Brownian motion defines a probability measure on the space $C[0,\infty)$ of continuous functions, called Wiener measure, which we denote by W.

A cylinder set $\mathcal{C}$ is a subset of $C[0,\infty)$ of the form

$$\mathcal{C} = \left\{B \in C[0,\infty) : B_{t_j} \in A_j \text{ for } 1 \le j \le n\right\} \tag{5.13}$$

where $0 < t_1 < \cdots < t_n$ and $A_1, \ldots, A_n$ are Borel subsets of $\mathbb{R}$. We may define $W : \mathcal{F} \to [0,1]$ as a probability measure on the σ-algebra $\mathcal{F}$ on $C[0,\infty)$ that is generated by the cylinder sets.

It follows from (5.8) that the Wiener measure of the set (5.13) is given by

$$W\{\mathcal{C}\} = C_n \int_A \exp\left[-\frac{1}{2}\left\{\frac{(x_n - x_{n-1})^2}{(t_n - t_{n-1})} + \cdots + \frac{(x_1 - x_0)^2}{(t_1 - t_0)}\right\}\right] dx_1 dx_2 \ldots dx_n$$

where $A = A_1 \times A_2 \times \cdots \times A_n \subset \mathbb{R}^n$, $x_0 = 0$, $t_0 = 0$, and

$$C_n = \frac{1}{\sqrt{2\pi(t_n - t_{n-1})\ldots(t_1 - t_0))}}.$$

If we suppose, for simplicity, that $t_i - t_{i-1} = \Delta t$, then we may write this expression as

$$W\{\mathcal{C}\} = C_n \int_A \exp\left[-\frac{\Delta t}{2}\left\{\left(\frac{x_n - x_{n-1}}{\Delta t}\right)^2 + \cdots + \left(\frac{x_1 - x_0}{\Delta t}\right)^2\right\}\right] dx_1 dx_2 \ldots dx_n$$

Thus, formally taking the limit as $n \to \infty$, we get the expression given in (3.89)

$$dW = C \exp\left[-\frac{1}{2}\int_0^t \dot{x}^2(s)\, ds\right] Dx \tag{5.14}$$

for the density of Wiener measure with respect to the (unfortunately nonexistent) 'flat' measure Dx. Note that, since Wiener measure is supported on the set of continuous functions that are nowhere differentiable, the exponential factor in (5.14) makes no more sense than the 'flat' measure.

It is possible interpret (5.14) as defining a Gaussian measure in an infinite dimensional Hilbert space, but we will not consider that theory here. Instead, we will describe some properties of Wiener measure suggested by (5.14) that are, in fact, true despite the formal nature of the expression.

First, as we saw in Section 14.3, Kac's version of the Feynman-Kac formula is suggested by (5.14). Although it is difficult to make sense of Feynman's expression for solutions of the Schrödinger equation as an oscillatory path integral, Kac's formula for the heat equation with a potential makes perfect sense as an integral with respect to Wiener measure.

Second, (5.14) suggests the *Cameron-Martin theorem*, which states that the translation $x(t) \mapsto x(t) + h(t)$ maps Wiener measure W to a measure W_h that is absolutely continuous with respect to Wiener measure if and only if $h \in H^1(0,t)$ has a square integrable derivative. A formal calculation based on (5.14), and the idea that, like Lebesgue measure, Dx should be invariant under translations gives

$$\begin{aligned} dW_h &= C \exp\left[-\frac{1}{2}\int_0^t \left\{\dot{x}(s) - \dot{h}(s)\right\}^2 \, ds\right] \, Dx \\ &= C \exp\left[\int_0^t \dot{x}(s)\dot{h}(s)\, ds - \frac{1}{2}\int_0^t \dot{h}^2(s)\, ds\right] \exp\left[-\frac{1}{2}\int_0^t \dot{x}^2(s)\, ds\right] \, Dx \\ &= \exp\left[\int_0^t \dot{x}(s)\dot{h}(s)\, ds - \frac{1}{2}\int_0^t \dot{h}^2(s)\, ds\right] \, dW. \end{aligned}$$

The integral

$$\langle x, h\rangle = \int_0^t \dot{x}(s)\dot{h}(s)\, ds = \int_0^t \dot{h}(s)\, dx(s)$$

may be defined as a Payley-Wiener-Zygmund integral (5.47) for any $h \in H^1$. We then get the Cameron-Martin formula

$$dW_h = \exp\left[\langle x, h\rangle - \frac{1}{2}\int_0^t \dot{h}^2(s)\, ds\right] \, dW. \tag{5.15}$$

Despite the formal nature of the computation, the result is correct.

Thus, although Wiener measure is not translation invariant (which is impossible for probability measures on infinite-dimensional linear spaces) it is 'almost' translation invariant in the sense that translations in a dense set of directions $h \in H^1$ give measures that are mutually absolutely continuous. On the other hand, if one translates Wiener measure by a function $h \notin H^1$, one gets a measure that is singular with respect to the original Wiener measure, and which is supported on a set of paths with different continuity and variation properties.

These results reflect the fact that Gaussian measures on infinite-dimensional spaces are concentrated on a dense set of directions, unlike the picture we have of a finite dimensional Gaussian measure with an invertible covariance matrice (whose density is spread out over an ellipsoid in all direction).

4. Brownian motion with drift

Brownian motion is a basic building block for the construction of a large class of Markov processes with continuous sample paths, called diffusion processes.

In this section, we discuss diffusion processes that have the same 'noise' as standard Brownian motion, but differ from it by a mean 'drift.' These process are defined by a stochastic ordinary differential equation (SDE) of the form

$$\dot{X} = b(X) + \xi(t), \tag{5.16}$$

where $b : \mathbb{R} \to \mathbb{R}$ is a given smooth function and $\xi(t) = \dot{B}(t)$ is, formally, the time derivative of Brownian motion, or 'white noise.' Equation (5.16) may be thought of as describing either a Brownian motion $\dot{X} = \xi$ perturbed by a drift term $b(X)$, or a deterministic ODE $\dot{X} = b(X)$ perturbed by an additive noise.

We begin with a heuristic discussion of white noise, and then explain more precisely what meaning we give to (5.16).

4.1. White noise

Although Brownian paths are not differentiable pointwise, we may interpret their time derivative in a distributional sense to get a generalized stochastic process called white noise. We denote it by

$$\xi(t, \omega) = \dot{B}(t, \omega).$$

We also use the notation $\xi dt = dB$. The term 'white noise' arises from the spectral theory of stationary random processes, according to which white noise has a 'flat' power spectrum that is uniformly distributed over all frequencies (like white light). This can be observed from the Fourier representation of Brownian motion in (5.7), where a formal term-by-term differentiation yields a Fourier series all of whose coefficients are Gaussian random variables with same variance.

Since Brownian motion has Gaussian independent increments with mean zero, its time derivative is a Gaussian stochastic process with mean zero whose values at different times are independent. (See Figure 2.) As a result, we expect the SDE (5.16) to define a Markov process X. This process is not Gaussian unless $b(X)$ is linear, since nonlinear functions of Gaussian variables are not Gaussian.

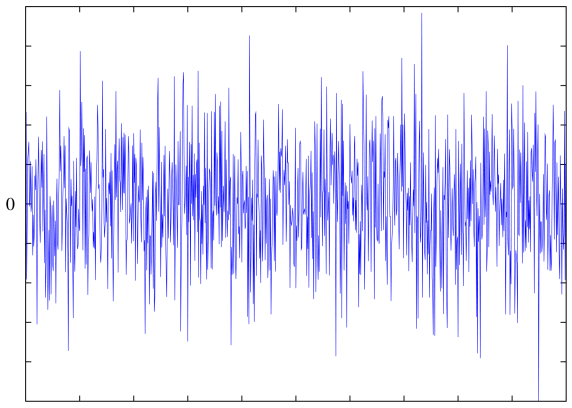

FIGURE 2. A numerical realization of an approximation to white noise.

To make this discussion more explicit, consider a finite difference approximation of ξ using a time interval of width Δt,

$$\xi_{\Delta t}(t) = \frac{B(t+\Delta t) - B(t)}{\Delta t}.$$

Then $\xi_{\Delta t}$ is a Gaussian stochastic process with mean zero and variance $1/\Delta t$. Using (5.11), we compute that its covariance is given by

$$\mathbf{E}\left[\xi_{\Delta t}(t)\xi_{\Delta t}(s)\right] = \delta_{\Delta t}(t-s)$$

where $\delta_{\Delta t}(t)$ is an approximation of the δ-function given by

$$\delta_{\Delta t}(t) = \frac{1}{\Delta t}\left(1 - \frac{|t|}{\Delta t}\right) \quad \text{if } |t| \leq \Delta t, \qquad \delta_{\Delta t}(t) = 0 \quad \text{otherwise.}$$

Thus, $\xi_{\Delta t}$ has a small but nonzero correlation time. Its power spectrum, which is the Fourier transform of its covariance, is therefore not flat, but decays at sufficiently high frequencies. We therefore sometimes refer to $\xi_{\Delta t}$ as 'colored noise.'

We may think of white noise ξ as the limit of this colored noise $\xi_{\Delta t}$ as $\Delta t \to 0$, namely as a δ-correlated stationary, Gaussian process with mean zero and covariance

$$\mathbf{E}\left[\xi(t)\xi(s)\right] = \delta(t-s). \tag{5.17}$$

In applications, the assumption of white noise is useful for modeling phenomena in which the correlation time of the noise is much shorter than any other time-scales of interest. For example, in the case of Brownian motion, the correlation time of the noise due to the impact of molecules on the Brownian particle is of the order of the collision time of the fluid molecules with each other. This is very small in comparison with the time-scales over which we use the SDE to model the motion of the particle.

4.2. Stochastic integral equations

While it is possible to define white noise as a distribution-valued stochastic process, we will not do so here. Instead, we will interpret white noise as a process whose time-integral is Brownian motion. Any differential equation that depends on white noise will be rewritten as an integral equation that depends on Brownian motion.

Thus, we rewrite (5.16) as the integral equation

$$X(t) = X(0) + \int_0^t b\left(X(s)\right)\, ds + B(t). \tag{5.18}$$

We use the differential notation

$$dX = b(X)dt + dB$$

as short-hand for the integral equation (5.18); it has no further meaning.

The standard Picard iteration from the theory of ODEs,

$$X_{n+1}(t) = X(0) + \int_0^t b\left(X_n(s)\right)\, ds + B(t),$$

implies that (5.18) has a unique continuous solution $X(t)$ for every continuous function $B(t)$, assuming that $b(x)$ is a Lipschitz-continuous function of x. Thus, if B is Brownian motion, the mapping $B(t) \mapsto X(t)$ obtained by solving (5.18) 'path by path' defines a stochastic process X with continuous sample paths. We call X a Brownian motion with drift.

Remark 5.16. According to Girsanov's theorem [**46**], the probability measure induced by X on $C[0,\infty)$ is absolutely continuous with respect to the Wiener measure induced by B, with density

$$\exp\left[\int_0^t b\left(X(s)\right)\, dX(s) - \frac{1}{2}\int_0^t b^2\left(X(s)\right)\, ds\right].$$

This is a result of the fact that the processes have the same 'noise,' so they are supported on the same paths; the drift changes only the probability density on those paths *c.f.* the Cameron-Martin formula (5.15).

4.3. The Fokker-Planck equation

We observed above that the transition density $p(x,t \mid y)$ of Brownian motion satisfies the diffusion equation (5.9). We will give a direct derivation of a generalization of this result for Brownian motion with drift.

We fix $y \in \mathbb{R}$ and write the conditional expectation given that $X(0) = y$ as

$$\mathbf{E}_y\left[\,\cdot\,\right] = \mathbf{E}\left[\,\cdot\,|X(0) = y\right].$$

Equation (5.18) defines a Markov process $X(t) = X_t$ with continuous paths. Moreover, as $\Delta t \to 0^+$, the increments of X satisfy

$$\mathbf{E}_y\left[X_{t+\Delta t} - X_t \mid X_t\right] = b\left(X_t\right)\Delta t + o\left(\Delta t\right), \tag{5.19}$$

$$\mathbf{E}_y\left[\left(X_{t+\Delta t} - X_t\right)^2 \mid X_t\right] = \Delta t + o\left(\Delta t\right), \tag{5.20}$$

$$\mathbf{E}_y\left[\left|X_{t+\Delta t} - X_t\right|^3 \mid X_t\right] = o(\Delta t), \tag{5.21}$$

where $o(\Delta t)$ denotes a term which approaches zero faster than Δt, meaning that

$$\lim_{\Delta t \to 0^+} \frac{o(\Delta t)}{\Delta t} = 0.$$

For example, to derive (5.19) we subtract (5.18) evaluated at $t + \Delta t$ from (5.18) evaluated at t to get

$$\Delta X = \int_t^{t+\Delta t} b\left(X_s\right)\, ds + \Delta B$$

where

$$\Delta X = X_{t+\Delta t} - X_t, \qquad \Delta B = B_{t+\Delta t} - B_t.$$

Using the smoothness of b and the continuity of X_t, we get

$$\begin{aligned}\Delta X &= \int_t^{t+\Delta t} \left[b\left(X_t\right) + o(1)\right]\, ds + \Delta B \\ &= b\left(X_t\right)\Delta t + \Delta B + o(\Delta t).\end{aligned}$$

Taking the expected value of this equation conditioned on X_t, using the fact that $\mathbf{E}[\Delta B] = 0$, and assuming we can exchange expectations with limits as $\Delta t \to 0^+$, we get (5.19). Similarly, Taylor expanding to second order, we find that the dominant term in $\mathbf{E}[(\Delta X)^2]$ is $\mathbf{E}[(\Delta B)^2] = \Delta t$, which gives (5.20). Equation (5.21) follows from the corresponding property of Brownian motion.

Now suppose that $\varphi : \mathbb{R} \to \mathbb{R}$ is any smooth test function with uniformly bounded derivatives, and let

$$e(t) = \frac{d}{dt}\mathbf{E}_y\left[\varphi\left(X_t\right)\right].$$

Expressing the expectation in terms of the transition density $p(x,t \mid y)$ of X_t, assuming that the time-derivative exists and that we may exchange the order of differentiation and expectation, we get

$$e(t) = \frac{d}{dt}\int \varphi(x)p(x,t \mid y)\,dx = \int \varphi(x)\frac{\partial p}{\partial t}(x,t \mid y)\,dx.$$

Alternatively, writing the time derivative as a limit of difference quotients, and Taylor expanding $\varphi(x)$ about $x = X_t$, we get

$$\begin{aligned} e(t) &= \lim_{\Delta t\to 0^+}\frac{1}{\Delta t}\mathbf{E}_y\left[\varphi\left(X_{t+\Delta t}\right)-\varphi\left(X_t\right)\right] \\ &= \lim_{\Delta t\to 0^+}\frac{1}{\Delta t}\mathbf{E}_y\left[\varphi'\left(X_t\right)\left(X_{t+\Delta t}-X_t\right)+\frac{1}{2}\varphi''\left(X_t\right)\left(X_{t+\Delta t}-X_t\right)^2+r_t(\Delta t)\right] \end{aligned}$$

where the remainder r_t satisfies

$$|r_t(\Delta t)| \le M\left|X_{t+\Delta t}-X_t\right|^3$$

for some constant M. Using the 'tower' property of conditional expectation (5.2) and (5.19), we have

$$\begin{aligned} \mathbf{E}_y\left[\varphi'\left(X_t\right)\left(X_{t+\Delta t}-X_t\right)\right] &= \mathbf{E}_y\left[\,\mathbf{E}_y\left[\varphi'\left(X_t\right)\left(X_{t+\Delta t}-X_t\right)\mid X_t\right]\right] \\ &= \mathbf{E}_y\left[\,\varphi'\left(X_t\right)\mathbf{E}_y\left[X_{t+\Delta t}-X_t\mid X_t\right]\right] \\ &= \mathbf{E}_y\left[\varphi'\left(X_t\right)b\left(X_t\right)\right]\Delta t. \end{aligned}$$

Similarly

$$\mathbf{E}_y\left[\,\varphi''\left(X_t\right)\left(X_{t+\Delta t}-X_t\right)^2\right] = \mathbf{E}_y\left[\,\varphi''\left(X_t\right)\right]\Delta t.$$

Hence,

$$e(t) = \mathbf{E}_y\left[\varphi'\left(X_t\right)b\left(X_t\right)+\frac{1}{2}\varphi''\left(X_t\right)\right].$$

Rewriting this expression in terms of the transition density, we get

$$e(t) = \int_{\mathbb{R}}\left[\varphi'(x)b(x)+\frac{1}{2}\varphi''(x)\right]p\left(x,t \mid y\right)\,dx.$$

Equating the two different expressions for $e(t)$ we find that,

$$\int_{\mathbb{R}}\varphi(x)\frac{\partial p}{\partial t}(x,t \mid y)\,dx = \int_{\mathbb{R}}\left[\varphi'(x)b(x)+\frac{1}{2}\varphi''(x)\right]p\left(x,t \mid y\right)\,dx.$$

This is the weak form of an advection-diffusion equation for the transition density $p(x,t \mid y)$ as a function of (x,t). After integrating by parts with respect to x, we find that, since φ is an arbitrary test function, smooth solutions p satisfy

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x}\left(bp\right)+\frac{1}{2}\frac{\partial^2 p}{\partial x^2}. \tag{5.22}$$

This PDE is called the Fokker-Planck, or forward Kolmogorov equation, for the diffusion process of Brownian motion with drift. When $b = 0$, we recover (5.9).

5. The Langevin equation

> A particle such as the one we are considering, large relative to the average distance between the molecules of the liquid and moving with respect to the latter at the speed ξ, experiences (according to Stokes' formula) a viscous resistance equal to $-6\pi\mu a\xi$. In actual fact, this value is only a mean, and by reason of the irregularity of the impacts of the surrounding molecules, the action of the fluid on the particle oscillates around the preceding value, to the effect that the equation of motion in the direction x is
>
> $$m\frac{d^2x}{dt^2} = -6\pi\mu a\frac{dx}{dt} + X.$$
>
> We know that the complementary force X is indifferently positive and negative and that its magnitude is such as to maintain the agitation of the particle, which, given the viscous resistance, would stop without it.[7]

In this section, we describe a one-dimensional model for the motion of a Brownian particle due to Langevin. A three-dimensional model may be obtained from the one-dimensional model by assuming that a spherical particle moves independently in each direction. For non-spherical particles, such as the pollen grains observed by Brown, rotational Brownian motion also occurs.

Suppose that a particle of mass m moves along a line, and is subject to two forces: (a) a frictional force that is proportional to its velocity; (b) a random white noise force. The first force models the average force exerted by a viscous fluid on a small particle moving though it; the second force models the fluctuations in the force about its mean value due to the impact of the fluid molecules.

This division could be questioned on the grounds that *all* of the forces on the particle, including the viscous force, ultimately arise from molecular impacts. One is then led to the question of how to derive a mesoscopic stochastic model from a more detailed kinetic model. Here, we will take the division of the force into a deterministic mean, given by macroscopic continuum laws, and a random fluctuating part as a basic hypothesis of the model. See Keizer [**30**] for further discussion of such questions.

We denote the velocity of the particle at time t by $V(t)$. Note that we consider the particle velocity here, not its position. We will consider the behavior of the position of the particle in Section 6. According to Newton's second law, the velocity satisfies the ODE

$$m\dot{V} = -\beta V + \gamma\xi(t), \tag{5.23}$$

where $\xi = \dot{B}$ is white noise, $\beta > 0$ is a damping constant, and γ is a constant that describes the strength of the noise. Dividing the equation by m, we get

$$\dot{V} = -bV + c\xi(t), \tag{5.24}$$

where $b = \beta/m > 0$ and $c = \gamma/m$ are constants. The parameter b is an inverse-time, so $[b] = T^{-1}$. Standard Brownian motion has dimension $T^{1/2}$ since $\mathbf{E}\left[B^2(t)\right] = t$, so white noise ξ has dimension $T^{-1/2}$, and therefore $[c] = LT^{-3/2}$.

---

[7] P. Langevin, *Comptes rendus Acad. Sci.* **146** (1908).

We suppose that the initial velocity of the particle is given by

$$V(0) = v_0, \tag{5.25}$$

where v_0 is a fixed deterministic quantity. We can obtain the solution for random initial data that is independent of the future evolution of the process by conditioning with respect to the initial value.

Equation (5.24) is called the *Langevin equation*. It describes the effect of noise on a scalar linear ODE whose solutions decay exponentially to the globally asymptotically stable equilibrium $V = 0$ in the absence of noise. Thus, it provides a basic model for the effect of noise on any system with an asymptotically stable equilibrium.

As explained in Section 4.2, we interpret (5.24)–(5.25) as an integral equation

$$V(t) = v_0 - b \int_0^t V(s)\, ds + cB(t), \tag{5.26}$$

which we write in differential notation as

$$dV = -bV dt + c dB.$$

The process $V(t)$ defined by (5.26) is called the *Ornstein-Uhlenbeck process*, or the OU process, for short.

We will solve this problem in a number of different ways, which illustrate different methods. In doing so, it is often convenient to use the formal properties of white noise; the correctness of any results we derive in this way can be verified directly.

One of the most important features of the solution is that, as $t \to \infty$, the process approaches a stationary process, called the stationary Ornstein-Uhlenbeck process. This corresponds physically to the approach of the Brownian particle to thermodynamic equilibrium in which the fluctuations caused by the noise balance the dissipation due to the damping terms. We will discuss the stationary OU process further in Section 6.

5.1. Averaging the equation

Since (5.24) is a linear equation for $V(t)$ with deterministic coefficients and an additive Gaussian forcing, the solution is also Gaussian. It is therefore determined by its mean and covariance. In this section, we compute these quantities by averaging the equation.

Let

$$\mu(t) = \mathbf{E}\left[V(t)\right].$$

Then, taking the expected value of (5.24), and using the fact that $\xi(t)$ has zero mean, we get

$$\dot{\mu} = -b\mu. \tag{5.27}$$

From (5.25), we have $\mu(0) = v_0$, so

$$\mu(t) = v_0 e^{-bt}. \tag{5.28}$$

Thus, the mean value of the process decays to zero in exactly the same way as the solution of the deterministic, damped ODE, $\dot{V} = -bV$.

Next, let

$$R\left(t, s\right) = \mathbf{E}\left[\left\{V(t) - \mu(t)\right\}\left\{V(s) - \mu(s)\right\}\right] \tag{5.29}$$

denote the covariance of the OU process. Then, assuming we may exchange the order of time-derivatives and expectations, and using (5.24) and (5.27), we compute that

$$\begin{aligned}\frac{\partial^2 R}{\partial t \partial s}(t,s) &= \mathbf{E}\left[\left\{\dot{V}(t) - \dot{\mu}(t)\right\}\left\{\dot{V}(s) - \dot{\mu}(s)\right\}\right] \\ &= \mathbf{E}\left[\{-b\left[V(t) - \mu(t)\right] + c\xi(t)\}\{-b\left[V(s) - \mu(s)\right] + c\xi(s)\}\right].\end{aligned}$$

Expanding the expectation in this equation and using (5.17), (5.29), we get

$$\frac{\partial^2 R}{\partial t \partial s} = b^2 R - bc\left\{L(t,s) + L(s,t)\right\} + c^2\delta(t-s) \tag{5.30}$$

where

$$L(t,s) = \mathbf{E}\left[\{V(t) - \mu(t)\}\xi(s)\right].$$

Thus, we also need to derive an equation for L. Note that $L(t,s)$ need not vanish when $t > s$ since then $V(t)$ depends on $\xi(s)$.

Using (5.24), (5.27), and (5.17), we find that

$$\begin{aligned}\frac{\partial L}{\partial t}(t,s) &= \mathbf{E}\left[\left\{\dot{V(t)} - \dot{\mu}(t)\right\}\xi(s)\right] \\ &= -b\mathbf{E}\left[\{V(t) - \mu(t)\}\xi(s)\right] + c\mathbf{E}\left[\xi(t)\xi(s)\right] \\ &= -bL(t,s) + c\delta(t-s).\end{aligned}$$

From the initial condition (5.25), we have

$$L(0,s) = 0 \qquad \text{for } s > 0.$$

The solution of this equation is

$$L(t,s) = \begin{cases} ce^{-b(t-s)} & \text{for } t > s, \\ 0 & \text{for } t < s. \end{cases} \tag{5.31}$$

This function solves the homogeneous equation for $t \neq s$, and jumps by c as t increases across $t = s$.

Using (5.31) in (5.30), we find that $R(t,s)$ satisfies the PDE

$$\frac{\partial^2 R}{\partial t \partial s} = b^2 R - bc^2 e^{-b|t-s|} + c^2\delta(t-s). \tag{5.32}$$

From the initial condition (5.25), we have

$$R(t,0) = 0, \quad R(0,s) = 0 \qquad \text{for } t, s > 0. \tag{5.33}$$

The second-order derivatives in (5.32) are the one-dimensional wave operator written in characteristic coordinates (t,s). Thus, (5.32)–(5.33) is a characteristic initial value problem for $R(t,s)$.

This problem has a simple explicit solution. To find it, we first look for a particular solution of the nonhomogeneous PDE (5.32). We observe that, since

$$\frac{\partial^2}{\partial t \partial s}\left(e^{-b|t-s|}\right) = -b^2 e^{-b|t-s|} + 2b\delta(t-s),$$

a solution is given by

$$R_p(t,s) = \frac{c^2}{2b}e^{-b|t-s|}.$$

Then, writing $R = R_p + \tilde{R}$, we find that $\tilde{R}(t,s)$ satisfies

$$\frac{\partial^2 \tilde{R}}{\partial t \partial s} = b^2\tilde{R}, \qquad \tilde{R}(t,0) = -\frac{c^2}{2b}e^{-bt}, \quad \tilde{R}(0,s) = -\frac{c^2}{2b}e^{-bs}.$$

This equation has the solution

$$\tilde{R}(t,s) = -\frac{c^2}{2b}\, e^{-b(t+s)}.$$

Thus, the covariance function (5.29) of the OU process defined by (5.24)–(5.25) is given by

$$R(t,s) = \frac{c^2}{2b}\left(e^{-b|t-s|} - e^{-b(t+s)}\right). \tag{5.34}$$

In particular, the variance of the process,

$$\sigma^2(t) = \mathbf{E}\left[\{V(t) - \mu(t)\}^2\right],$$

or $\sigma^2(t) = R(t,t)$, is given by

$$\sigma^2(t) = \frac{c^2}{2b}\left(1 - e^{-2bt}\right), \tag{5.35}$$

and the one-point probability density of the OU process is given by the Gaussian density

$$p(v,t) = \frac{1}{\sqrt{2\pi\sigma^2(t)}} \exp\left\{-\frac{[v-\mu(t)]^2}{2\sigma^2(t)}\right\}. \tag{5.36}$$

The success of the method used in this section depends on the fact that the Langevin equation is linear with additive noise. For nonlinear equations, or equations with multiplicative noise, one typically encounters the 'closure' problem, in which higher order moments appear in equations for lower order moments, leading to an infinite system of coupled equations for averaged quantities. In some problems, it may be possible to use a (more or less well-founded) approximation to truncate this infinite system to a finite system.

5.2. Exact solution

The SDE (5.24) is sufficiently simple that we can solve it exactly. A formal solution of (5.24) is

$$V(t) = v_0 e^{-bt} + c\int_0^t e^{-b(t-s)}\xi(s)\, ds. \tag{5.37}$$

Setting $\xi = \dot{B}$, and using a formal integration by parts, we may rewrite (5.37) as

$$V(t) = v_0 e^{-bt} + bc\left(B(t) - \int_0^t e^{-b(t-s)} B(s)\, ds\right). \tag{5.38}$$

This last expression does not involve any derivatives of $B(t)$, so it defines a continuous function $V(t)$ for any continuous Brownian sample function $B(t)$. One can verify by direct calculation that (5.38) is the solution of (5.26).

The random variable $V(t)$ defined by (5.38) is Gaussian. Its mean and covariance may be computed most easily from the formal expression (5.37), and they agree with the results of the previous section.

For example, using (5.37) in (5.29) and simplifying the result by the use of (5.17), we find that the covariance function is

$$\begin{aligned} R(t,s) &= c^2 \mathbf{E}\left[\left\{\int_0^t e^{-b\left(t-t'\right)}\xi\left(t'\right)\,dt'\right\}\left\{\int_0^s e^{-b\left(s-s'\right)}\xi\left(s'\right)\,ds'\right\}\right] \\ &= c^2\int_0^t\int_0^s e^{-b\left(t+s-t'-s'\right)}\mathbf{E}\left[\xi\left(t'\right)\xi\left(s'\right)\right]\,ds'dt' \\ &= c^2\int_0^t\int_0^s e^{-b\left(t+s-t'-s'\right)}\delta\left(t'-s'\right)\,ds'dt' \\ &= \frac{c^2}{2b}\left\{e^{-b|t-s|}-e^{-b(t+s)}\right\}. \end{aligned}$$

In more complicated problems, it is typically not possible to solve a stochastic equation exactly for each realization of the random coefficients that appear in it, so we cannot compute the statistical properties of the solution by averaging the exact solution. We may, however, be able to use perturbation methods or numerical simulations to obtain approximate solutions whose averages can be computed.

5.3. The Fokker-Planck equation

The final method we use to solve the Langevin equation is based on the Fokker-Planck equation. This method depends on a powerful and general connection between diffusion processes and parabolic PDEs.

From (5.22), the transition density $p(v,t \mid w)$ of the Langevin equation (5.24) satisfies the diffusion equation

$$\frac{\partial p}{\partial t} = \frac{\partial}{\partial v}\left(bvp\right) + \frac{1}{2}c^2\frac{\partial^2 p}{\partial v^2}. \tag{5.39}$$

Note that the coefficient of the diffusion term is proportional to c^2 since the Brownian motion cB associated with the white noise $c\xi$ has quadratic variation $\mathbf{E}\left[(c\Delta B)^2\right] = c^2\Delta t$.

To solve (5.39), we write it in characteristic coordinates associated with the advection term. (An alternative method is to Fourier transform the equation with respect to v, which leads to a first-order PDE for the transform since the variable coefficient term involves only multiplication by v. This PDE can then be solved by the method of characteristics.)

The sub-characteristics of (5.39) are defined by

$$\frac{dv}{dt} = -bv,$$

whose solution is $v = \tilde{v}e^{-bt}$. Making the change of variables $v \mapsto \tilde{v}$ in (5.39), we get

$$\frac{\partial p}{\partial t} = bp + \frac{1}{2}c^2 e^{2bt}\frac{\partial^2 p}{\partial \tilde{v}^2},$$

which we may write as

$$\frac{\partial}{\partial t}\left(e^{-bt}p\right) = \frac{1}{2}c^2 e^{2bt}\frac{\partial^2}{\partial \tilde{v}^2}\left(e^{-bt}p\right).$$

To simplify this equation further, we define

$$\tilde{p} = e^{-bt}p, \qquad \tilde{t} = \frac{c^2}{2b}\left(e^{2bt}-1\right),$$

which gives the standard diffusion equation

$$\frac{\partial \tilde{p}}{\partial \tilde{t}} = \frac{1}{2}\frac{\partial^2 \tilde{p}}{\partial \tilde{v}^2}.$$

The solution with initial condition

$$\tilde{p}\left(\tilde{v}, 0\right) = \delta\left(\tilde{v} - v_0\right)$$

is given by

$$\tilde{p}\left(\tilde{v}, \tilde{t}\right) = \frac{1}{(2\pi\tilde{t})^{1/2}}\, e^{-(\tilde{v}-v_0)^2/(2\tilde{t})}.$$

Rewriting this expression in terms of the original variables, we get (5.36).

The corresponding expression for the transition density is

$$p\left(v, t \mid v_0\right) = \frac{1}{\sqrt{2\pi\sigma^2(t)}} \exp\left\{-\frac{\left[v - v_0 e^{-bt}\right]^2}{2\sigma^2(t)}\right\}$$

where σ is given in (5.35).

Remark 5.17. It is interesting to note that the Ornstein-Uhlenbeck process is closely related to the 'imaginary' time version of the quantum mechanical simple harmonic oscillator. The change of variable

$$p(v,t) = \exp\left(\frac{1}{2}bx^2 - bt\right)\psi(x,t) \qquad v = cx,$$

transforms (5.39) to the diffusion equation with a quadratic potential

$$\frac{\partial \psi}{\partial t} = \frac{1}{2}\frac{\partial^2 \psi}{\partial x^2} - \frac{1}{2}b^2x^2\psi.$$

6. The stationary Ornstein-Uhlenbeck process

As $t \to \infty$, the Ornstein-Uhlenbeck process approaches a stationary Gaussian process with zero mean, called the stationary Ornstein-Uhlenbeck process. This approach occurs on a time-scale of the order b^{-1}, which is the time-scale for solutions of the deterministic equation $\dot{V} = -bV$ to decay to zero.

From (5.36) and (5.35), the limiting probability density for v is a Maxwellian distribution,

$$p(v) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-v^2/(2\sigma^2)} \tag{5.40}$$

with variance

$$\sigma^2 = \frac{c^2}{2b}. \tag{5.41}$$

We can also obtain (5.40) by solving the ODE for steady solutions of (5.39)

$$\frac{1}{2}c^2\frac{d^2p}{dv^2} + b\frac{d}{dv}\left(vp\right) = 0.$$

Unlike Brownian paths, whose fluctatations grow with time, the stationary OU paths consist of fluctuations that are typically of the order σ, although larger fluctuations occur over long enough times.

The stationary OU process is the exact solution of the SDE (5.24) if, instead of taking deterministic initial conditions, we suppose that $V(0)$ is a Gaussian random variable with the stationary distribution (5.40).

Taking the limit as $t \to \infty$ in (5.34), we find that the covariance function of the stationary OU process is

$$R(t-s) = \sigma^2 \, e^{-b|t-s|}. \tag{5.42}$$

The covariance function depends only on the time-difference since the process is stationary. Equation (5.42) shows that the values of the stationary OU process become uncorrelated on the damping time-scale b^{-1}.

6.1. Parameter values for Brownian motion

Before we use use the OU process to determine the spatial diffusion of a Brownian particle, we give some typical experimental parameters for Brownian motion [**38**] and discuss their implications.

A typical radius of a spherical Brownian particle in water (for example, a polystyrene microsphere) is $a = 10^{-6}\,\mathrm{m}$. Assuming that the density of the particle is close to the density of water, its mass is approximately $m = 4 \times 10^{-15}\,\mathrm{Kg}$. According to Stokes law (2.24), at low Reynolds numbers, the viscous drag on a sphere of radius a moving with velocity v through a fluid with viscosity μ is equal to $6\pi\mu a v$. Thus, in (5.23), we take

$$\beta = 6\pi\mu a.$$

The viscosity of water at standard conditions is approximately $\mu = 10^{-3}\,\mathrm{Kg\,m^{-1}s^{-1}}$, which gives $\beta = 2 \times 10^{-8}\,\mathrm{Kg\,s^{-1}}$.

The first conclusion from these figures is that the damping time,

$$\frac{1}{b} = \frac{m}{\beta} \approx 2 \times 10^{-7}\,\mathrm{s},$$

is very small compared with the observation times of Brownian motion, which are typically on the order of seconds. Thus, we can assume that the Brownian particle velocity is in thermodynamic equilibrium and is distributed according to the stationary OU distribution. It also follows that the stationary OU fluctuations are very fast compared with the time scales of observation.

Although b^{-1} is small compared with macroscopic time-scales, it is large compared with molecular time scales; the time for water molecules to collide with each other is of the order of $10^{-11}\,\mathrm{s}$ or less. Thus, it is appropriate to use white noise to model the effect of fluctuations in the molecular impacts.

We can determine the strength of the noise in (5.23) by an indirect argument. According to statistical mechanics, the equilibrium probability density of a Brownian particle is proportional to $\exp(-E/kT)$, where $E = \frac{1}{2}mv^2$ is the kinetic energy of the particle, k is Boltzmann's constant, and T is the absolute temperature. This agrees with (5.40) if

$$\sigma^2 = \frac{kT}{m}. \tag{5.43}$$

At standard conditions, we have $kT = 4 \times 10^{-21}\,\mathrm{J}$, which gives $\sigma = 10^{-3}\,\mathrm{ms^{-1}}$. This is the order of magnitude of the thermal velocity fluctuations of the particle. The corresponding Reynolds numbers $\mathrm{R} = Ua/\nu$ are of the order 10^{-3} which is consistent with the use of Stokes' law.

Remark 5.18. It follows from (5.41) and (5.43) that

$$\gamma^2 = 2kT\beta.$$

This equation is an example of a *fluctuation-dissipation theorem*. It relates the macroscopic damping coefficient β in (5.23) to the strength γ^2 of the fluctuations when the system in thermodynamic equilibrium at temperature T.

6.2. The spatial diffusion of Brownian particles

Let us apply these results to the spatial diffusion of Brownian particles. We assume that the particles are sufficiently dilute that we can neglect any interactions between them.

Let $X(t)$ be the position at time t of a particle in Brownian motion measured along some coordinate axis. We assume that its velocity $V(t)$ satisfies the Langevin equation (5.23). Having solved for $V(t)$, we can obtain X by an integration

$$X(t) = \int_0^t V(s)\, ds.$$

Since $X(t)$ is a linear function of the Gaussian process $V(t)$, it is also Gaussian. The stochastic properties of X may be determined exactly from those of V, for example by averaging this equation to find its mean and covariance. We can, however, simplify the calculation when the parameters have the order of magnitude of the experimental ones given above.

On the time-scales over which we want to observe $X(t)$, the velocity $V(t)$ is a rapidly fluctuating, stationary Gaussian process with zero mean and a very short correlation time b^{-1}. We may therefore approximate it by white noise. From (5.42), the covariance function $R(t-s) = \mathbf{E}\,[V(t)V(s)]$ of V is given by

$$\mathbf{E}\,[V(t)V(s)] = \frac{2\sigma^2}{b}\left(\frac{be^{-b|t-s|}}{2}\right)$$

As $b \to \infty$, we have $be^{-b|t|}/2 \rightharpoonup \delta(t)$. Thus, from (5.17), if $bt \gg 1$, we may make the approximation

$$V(t) = \sqrt{\frac{2\sigma^2}{b}}\,\xi(t)$$

where $\xi(t)$ is a standard white noise.

It then follows that the integral of $V(t)$ is given in terms of a standard Brownian motion $B(t)$ by

$$X(t) = \sqrt{\frac{2\sigma^2}{b}}\,B(t).$$

The probability distribution of $X(t)$, which we denote $p(x,t)$, therefore satisfies the diffusion equation

$$\frac{\partial p}{\partial t} = D\frac{\partial^2 p}{\partial^2 x}$$

where, $D = \sigma^2/b$, or by use of (5.43),

$$D = \frac{kT}{\beta} \tag{5.44}$$

This is the result derived by Einstein (1905).

As Einstein observed, one can use (5.44) to determine Avogadro's number N_A, the number of molecules in one mole of gas, by measuring the diffusivity of Brownian

particles. Boltzmann's constant k is related to the macroscopically measurable gas constant R by $R = kN_A$; at standard conditions, we have $RT \approx 2,400\,\mathrm{J}$. Thus,

$$N_A = \frac{RT}{\beta D}$$

For the experimental values given above, with $\beta = 2 \times 10^{-8}\,\mathrm{Kg\,s^{-1}}$, the diffusivity of Brownian particles is found to be approximately $2 \times 10^{-13}\,\mathrm{m^2\,s^{-1}}$, meaning that the particles diffuse a distance on the order of a micron over a few seconds [**38**]. This gives $N_A \approx 6 \times 10^{23}$, consistent with the accepted value of $N_A = 6.02214 \times 10^{23}$, measured more accurately by other methods.

7. Stochastic differential equations

In this section, we discuss SDEs that are driven by white noise whose strength depends on the solution. Our aim here is to introduce some of the main ideas, rather than give a full discussion, and we continue to consider scalar SDEs. The ideas generalize to systems of SDEs, as we briefly explain in Section 7.5. For a more detailed introduction to the theory of SDEs, see [**19**]. For the numerical solution of SDEs, see [**32**]

The SDE (5.16) considered in Section 4 contains white noise with a constant strength. If the strength of the white noise depends on the solution, we get an SDE of the form

$$\dot{X} = b(X,t) + \sigma(X,t)\xi(t), \tag{5.45}$$

where $b, \sigma : \mathbb{R} \times [0,\infty) \to \mathbb{R}$ are smooth coefficient functions, which describe the drift and diffusion, respectively. We allow the coefficients to depend explicitly on t.

As we will explain, there is a fundamental ambiguity in how to interpret an SDE such as (5.45) which does not arise when σ is constant.

First, we rewrite (5.45) as an integral equation for $X(t)$,

$$X(t) = X(0) + \int_0^t b\left(X(s), s\right)\, ds + \int_0^t \sigma\left(X(s), s\right) dB(s), \tag{5.46}$$

or, in differential notation, as

$$dX = b(X,t)\, dt + \sigma(X,t)\, dB.$$

We interpret (5.45) as the corresponding integral equation (5.46). In order to do so, we need to define the stochastic integral

$$\int_0^t \sigma\left(X(s), s\right) dB(s).$$

When $\sigma = 1$, we made the obvious definition that this integral is to equal $B(t)$. More generally, if $F(t)$ is a stochastic process with smooth sample paths, we can define the integral of F against dB by use of a formal integration by parts:

$$\int_0^t F(s) dB(s) = F(t)B(t) - \int_0^t \dot{F}(s)B(s)\, ds.$$

For deterministic integrands, we can relax the smoothness condition and define a stochastic integral for any $f \in L^2(0,t)$ such that

$$\int_0^t f^2(s)\, ds < \infty.$$

If $f(s)$ is smooth and $f(t) = 0$, then (by a formal white-noise computation, which is easy to verify [**19**])

$$\begin{aligned}\mathbf{E}\left[\left\{\int_0^t f(s)dB(s)\right\}^2\right] &= \int_0^t \int_0^t f(s)f(r)\mathbf{E}\left[\xi(s)\xi(r)\right]\, dsdr \\ &= \int_0^t \int_0^t f(s)f(r)\delta(s-r)\, dsdr \\ &= \int_0^t f^2(s)\, ds.\end{aligned}$$

If $f \in L^2(0,t)$, then we choose a sequence of smooth functions f_n such that $f_n \to f$ with respect to the L^2-norm. We then define

$$\int_0^t f(s)dB(s) = \lim_{n\to\infty} \int_0^t f_n(s)dB(s), \tag{5.47}$$

where, from the preceding estimate, the integrals converge in the sense of mean-square expectation,

$$\lim_{n\to\infty} \mathbf{E}\left[\left\{\int_0^t f_n(s)dB(s) - \int_0^t f(s)dB(s)\right\}^2\right] = 0.$$

This definition of a stochastic integral is due to Payley, Wiener, and Zygmund (1933).

None of these definitions work, however, if F is a stochastic process with continuous but non-differentiable paths, such as a function of B or X of (5.46), which is exactly the case we are interested in.

In the next section, we illustrate the difficulties that arise for such integrals. We will then indicate how to define the Itô integral, which includes the above definitions as special cases.

7.1. An illustrative stochastic integral

Let $B(t)$ be a standard Brownian motion starting at 0. Consider, as a specific example, the question of how to define the integral

$$J(t) = \int_0^t B(s)\, dB(s) \tag{5.48}$$

by the use of Riemann sums. We will give two different definitions, corresponding to the Strantonovich and Itô integral, respectively.

Let $0 = s_0 < s_1 < \cdots < s_n < s_{n+1} = t$ be a non-random partition of $[0,t]$. The Strantonovich definition of (5.48) corresponds to a limit of centered Riemann sums, such as

$$J_n^{(\mathrm{S})} = \sum_{i=0}^{n} \frac{1}{2}\left[B\left(s_{i+1}\right) + B\left(s_i\right)\right]\left[B\left(s_{i+1}\right) - B\left(s_i\right)\right].$$

This gives a telescoping series with the sum

$$J_n^{(\mathrm{S})} = \frac{1}{2}\sum_{i=0}^{n}\left[B^2\left(s_{i+1}\right) - B^2\left(s_i\right)\right] = \frac{1}{2}\left[B^2\left(s_{n+1}\right) - B^2\left(s_0\right)\right].$$

Thus, we get the Strantonovich integral

$$\int_0^{t\,(\mathrm{S})} B(s)\,dB(s) = \frac{1}{2}B^2(t), \tag{5.49}$$

as in the usual calculus. The Strantonovich definition of the integral is, however, not well-suited to the Markov and martingale properties of stochastic processes. For example, the expected value of the Strantonovich integral in (5.49) is nonzero and equal to $t/2$.

The Itô definition of (5.48) corresponds to a limit of forward-differenced Riemann sums, such as

$$J_n^{(\mathrm{I})} = \sum_{i=0}^{n} B\left(s_i\right)\left[B\left(s_{i+1}\right) - B\left(s_i\right)\right].$$

We can rewrite this equation as

$$\begin{aligned} J_n^{(\mathrm{I})} &= \frac{1}{2}\sum_{i=0}^{n}\left[\left\{B\left(s_{i+1}\right) + B\left(s_i\right)\right\} - \left\{B\left(s_{i+1}\right) - B_i\left(s_i\right)\right\}\right]\left[B\left(s_{i+1}\right) - B\left(s_i\right)\right] \\ &= \frac{1}{2}\sum_{i=0}^{n}\left[B^2\left(s_{i+1}\right) - B^2\left(s_i\right)\right] - \frac{1}{2}\sum_{i=0}^{n}\left[B\left(s_{i+1}\right) - B\left(s_i\right)\right]^2. \end{aligned}$$

The first sum gives $B^2(t)$, as for the Strantonovich integral, while the second sum converges almost surely to t as $n \to \infty$ by the quadratic-variation property of Brownian motion.

The Itô integral is therefore

$$\int_0^{t\,(\mathrm{I})} B(s)\,dB(s) = \frac{1}{2}\left[B^2(t) - t\right]. \tag{5.50}$$

This definition has powerful stochastic properties; for example, it defines a martingale, consistent with the fact that the expected value of the Itô integral in (5.50) is equal to zero.

If we use the Itô definition, however, the usual rules of calculus must be modified to include (5.12). For example, the differential form of (5.50) may be derived formally as follows:

$$d\left(\frac{1}{2}B^2\right) = \frac{1}{2}\left[(B + dB)^2 - B^2\right] = BdB + \frac{1}{2}(dB)^2 = BdB + \frac{1}{2}dt. \tag{5.51}$$

As this example illustrates, there is an inherent ambiguity in how one defines stochastic integrals such as (5.48). This ambiguity is caused by the sensitivity of the values of the Riemann sums to the location of the point where one evaluates the integrand, which is a result of the unbounded total variation of the Brownian sample paths.

We will use the Itô definition, but it should be emphasized that this choice is a matter of mathematical convenience. For instance, one can express the Itô and Strantonovich integrals in terms of each other.

7.2. The Itô integral

We will not define the Itô's integral in detail, but we will give a brief summary of some of the main points. Evans [**19**] or Varadhan [**46**] give proofs of most of the results stated here.

A stochastic process

$$F : [0, \infty) \times \Omega \to \mathbb{R}$$

is said to be *adapted* to a Brownian motion $B(t)$ if, for each $t \geq 0$, $F(t)$ is measurable with respect to the σ-algebra $\mathcal{F}_t$ generated by the random variables $\{B(s) : 0 \leq s \leq t\}$. Roughly speaking, this means that $F(t)$ is a function of $\{B(s) : 0 \leq s \leq t\}$.

If $F(t)$ is an adapted process with almost surely continuous sample paths and

$$\int_0^t \mathbf{E}\left[F^2(s)\right]\, ds < \infty,$$

then we can define the stochastic Itô integral of F with respect to B as a limit in mean-square expectation of forward-differenced Riemann sums

$$\int_0^t F(s)\, dB(s) = \lim_{n\to\infty} \sum_{i=0}^{n} F\left(s_i\right)\left[B\left(s_{i+1}\right) - B\left(s_i\right)\right],$$

or, in general, as a limit of integrals of adapted simple functions.

An important property of the Itô integral is that, as in (5.50),

$$\mathbf{E}\left[\int_0^t F(s)\, dB(s)\right] = 0. \tag{5.52}$$

This follows because $F(t)$ is independent of $B(t + \Delta t) - B(t)$ for $\Delta t > 0$, since F is adapted, so

$$\mathbf{E}\left[F\left(s_i\right)\left\{B\left(s_{i+1}\right) - B\left(s_i\right)\right\}\right] = \mathbf{E}\left[F\left(s_i\right)\right]\mathbf{E}\left[B\left(s_{i+1}\right) - B\left(s_i\right)\right] = 0.$$

Since Brownian motion has independent increments, one can see by a similar argument that the Itô integral

$$M(t) = M_0 + \int_0^t F(s)\, dB(s) \tag{5.53}$$

defines a martingale, meaning that $\mathbf{E}\left[M(t) \mid \mathcal{F}_s\right] = M(s)$ for $0 \leq s < t$.

We then define the Itô SDE

$$dX = b(X, t)\, dt + \sigma(X, t)\, dB \tag{5.54}$$

by (5.46), where the integral is understood to be an Itô integral. The initial data

$$X(0) = X_0 \tag{5.55}$$

is a given $\mathcal{F}_0$-measurable random variable. Here, we allow the initial value $B(0)$ of the Brownian motion to be a random variable, and $\mathcal{F}_0 = \sigma\left(B(0)\right)$.

For the SDE (5.18) with constant noise, we can define solutions 'path by path.' For (5.46), the definition of a solution depends on a probabilistic convergence of the integral. Thus, it is essentially stochastic in nature.

It can be shown that the SDE (5.54)–(5.55) has a unique adapted solution $X(t)$ with continuous paths defined for all $0 \leq t \leq T$ if, for example:

(1) the functions $b, \sigma : \mathbb{R} \times [0, T] \to \mathbb{R}$ are continuous, globally Lipschitz in x, and uniformly bounded in t, meaning that there exists a constant K such that for all $x, y \in \mathbb{R}$, $t \in [0, T]$

$$|b(x,t) - b(y,t)| \leq K|x - y|, \quad |\sigma(x,t) - \sigma(y,t)| \leq K|x - y|,$$
$$|b(x,t)| \leq K\left(1 + |x|\right), \quad |\sigma(x,t)| \leq K\left(1 + |x|\right);$$

(2) the initial data satisfies

$$\mathbf{E}\left[X_0^2\right] < \infty. \tag{5.56}$$

Such solutions are called strong solutions. It is also possible to define weak solutions of that satisfy the SDE in a distributional sense, and which exist even if the coefficient functions are not Lipschitz continuous, but we will not use weak solutions here.

7.3. Itô's formula

As we saw above, it is necessary to modify the usual rules of calculus if one uses Itô integrals. The key result is a version of the chain rule called Itô's formula.

Suppose that $X(t)$ is a solution of the Itô SDE (5.54), and $f(X,t)$ is a smooth function $f : \mathbb{R} \times [0,\infty) \to \mathbb{R}$. Here, we abuse notation slightly and use the same symbol for the argument of f and the process. Define

$$Y(t) = f\left(X(t), t\right).$$

Then Itô's formula states that Y satisfies the SDE

$$\begin{aligned} dY = &\left[\frac{\partial f}{\partial t}(X,t) + b(X,t)\frac{\partial f}{\partial X}(X,t) + \frac{1}{2}\sigma^2(X,t)\frac{\partial^2 f}{\partial X^2}(X,t)\right] dt \\ &+ \sigma(X,t)\frac{\partial f}{\partial X}(X,t)\, dB. \end{aligned} \tag{5.57}$$

This equation stands, of course, for the corresponding stochastic integral equation. Equation (5.57) is what one would obtain from the usual chain rule with an additional term in the drift proportional to the second x-derivative of f. In particular, if $X = B$, then $b = 0$, $\sigma = 1$, and Itô's formula becomes

$$df(B,t) = \left[\frac{\partial f}{\partial t}(B,t) + \frac{1}{2}\frac{\partial f^2}{\partial B^2}(B,t)\right] dt + \frac{\partial f}{\partial B}(B,t)\, dB. \tag{5.58}$$

For a proof of (5.57), see [**19**].

Itô's formula may be motivated by a formal computation using (5.54) and (5.12). For example, when $Y = f(X)$ we get, denoting X-derivatives by primes,

$$\begin{aligned} dY &= f'(X)dX + \frac{1}{2}f''(X)dX^2 \\ &= f'(X)\left[b(X,t)\, dt + \sigma(X,t)\, dB\right] + \frac{1}{2}f''(X)\sigma^2(X,t)\ (dB)^2 \\ &= \left[f'(X)b(X,t) + \frac{1}{2}f''(X)\sigma^2(X,t)\right] dt + f'(X)\sigma(X,t)\, dB. \end{aligned}$$

Example 5.19. Itô's formula (5.58) gives, as in (5.51),

$$d\left(\frac{1}{2}B^2\right) = \frac{1}{2}dt + BdB.$$

Example 5.20. If $f(B) = e^{\sigma B}$, where σ is a constant, then (5.58) implies that

$$de^{\sigma B} = \frac{1}{2}\sigma^2 e^{\sigma B}\, dt + \sigma e^{\sigma B}\, dB.$$

Taking expected values of this equation, and using the martingale property (5.52) of the Itô integral, we find that

$$d\mathbf{E}\left[e^{\sigma B}\right] = \frac{1}{2}\sigma^2 \mathbf{E}\left[e^{\sigma B}\right].$$

Solving this equation, and assuming that $B(t)$ starts at 0, we find that

$$\mathbf{E}\left[e^{\sigma B}\right] = e^{\sigma^2 t/2}. \tag{5.59}$$

7.4. The Fokker-Planck equation

Itô's formula provides a quick and efficient way to derive the Fokker-Planck equation. Suppose that $X(t)$ satisfies

$$dX = b(X,t)\,dt + \sigma(X,t)\,dB.$$

Taking the expectation of Itô's formula (5.57) and using the martingale property (5.52), we find that for any smooth function $f : \mathbb{R} \to \mathbb{R}$,

$$\mathbf{E}\left[f\left(X(t)\right)\right] = \int_0^t \mathbf{E}\left[f'\left(X(s)\right) b\left(X(s),s\right) + \frac{1}{2} f''\left(X(s)\right) \sigma^2\left(X(s),s\right)\right] ds.$$

Differentiating this equation with respect to t, we get

$$\frac{d}{dt}\mathbf{E}\left[f\left(X(t)\right)\right] = \mathbf{E}\left[f'\left(X(t)\right) b\left(X(t),t\right) + \frac{1}{2} f''\left(X(t)\right) \sigma^2\left(X(t),t\right)\right].$$

Writing this equation in terms of the probability density $p(x,t)$, or the transition density $p(x,t \mid y,s)$ if we condition on $X(s) = y$, we get

$$\frac{d}{dt}\int f(x)\,p(x,t)\,dx = \int \left[f'(x)\,b(x,t) + \frac{1}{2} f''(x)\,\sigma^2(x,t)\right] p(x,t)\,dx,$$

which is the weak form of the Fokker-Planck equation,

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x}\left(b(x,t)\,p\right) + \frac{1}{2}\frac{\partial^2}{\partial^2 x}\left(\sigma^2(x,t)\,p\right). \tag{5.60}$$

7.5. Systems of SDEs

A system of SDEs for a vector-valued stochastic process $\vec{X}(t) = (X_1(t), \ldots, X_n(t))$ may be written as

$$d\vec{X} = \vec{b}\left(\vec{X},t\right)\,dt + \sigma\left(\vec{X},t\right)\,d\vec{B}. \tag{5.61}$$

In (5.61), the vector $\vec{B}(t) = (B_1(t), \ldots, B_n(t))$ is an n-dimensional Brownian motion whose components $B_i(t)$ are independent one-dimensional Brownian motions such that

$$\mathbf{E}\left[B_i(t)B_j(s)\right] = \begin{cases} \min(t,s) & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

The coefficient functions in (5.61) are a drift vector $\vec{b} = (b_1, \ldots, b_n)$ and a diffusion matrix $\sigma = (\sigma_{ij})$

$$\vec{b} : \mathbb{R}^n \times [0,\infty) \to \mathbb{R}^n, \qquad \sigma : \mathbb{R}^n \times [0,\infty) \to \mathbb{R}^{n\times n},$$

which we assume satisfy appropriate smoothness conditions.

The differential form of the SDE (5.61) is short-hand for the integral equation

$$\vec{X}(t) = \vec{X}_0 + \int_0^t \vec{b}\left(\vec{X}(s),s\right)\,ds + \int_0^t \sigma\left(\vec{X}(s),s\right)\,d\vec{B}(s),$$

or, in component form,

$$X_i(t) = X_{i0} + \int_0^t b_i\left(X_1(s), \ldots, X_n(s), s\right)\, ds + \sum_{j=1}^n \int_0^t \sigma_{ij}\left(X_1(s), \ldots, X_n(s), s\right)\, dB_j(s) \qquad \text{for } 1 \leq i \leq n.$$

The integrals here are understood as Itô integrals.

If $f : \mathbb{R}^n \times [0, \infty) \to \mathbb{R}$ is a smooth function $f(X_1, \ldots, X_n, t)$, and

$$Y(t) = f\left(X_1(t), \ldots, X_n(t), t\right)$$

where $\vec{X}(t) = (X_1(t), \ldots, X_n(t))$ is a solution of (5.61), then Itô's formula is

$$dY = \left(\frac{\partial f}{\partial t} + \sum_{i=1}^n b_i \frac{\partial f}{\partial X_i} + \frac{1}{2} \sum_{i,j,k=1}^n \sigma_{ik}\sigma_{jk} \frac{\partial^2 f}{\partial X_i \partial X_j} \right) dt + \sum_{i,j=1}^n \sigma_{ij} \frac{\partial f}{\partial X_i}\, dB_j. \tag{5.62}$$

This result follows formally from the generalization of (5.12) to the 'rule'

$$dB_i\, dB_j = \begin{cases} dt & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

The coefficients of the resulting drift terms in (5.62) are

$$a_{ij} = \sum_{k=1}^n \sigma_{ik}\sigma_{jk}.$$

Thus, $A = (a_{ij})$ is given by $A = \sigma\sigma^\top$.

The Fokker-Planck equation for the transition density $p\left(\vec{x}, t \mid \vec{y}, s\right)$ may be derived in the same way as in the scalar case. The result is that

$$\frac{\partial p}{\partial t} = -\sum_{i=1}^n \frac{\partial}{\partial x_i}\left(b_i p\right) + \frac{1}{2} \sum_{i,j=1}^n \frac{\partial^2}{\partial x_i \partial x_j}\left(a_{ij} p\right),$$

with the initial condition $p\left(\vec{x}, s \mid \vec{y}, s\right) = \delta\left(\vec{x} - \vec{y}\right)$.

7.6. Strantonovich SDEs

Suppose that X satisfies the Strantonovich SDE

$$dX = b(X, t)\, dt + \sigma(X, t) \circ dB \tag{5.63}$$

where the notation $\circ dB$ indicates that the corresponding integral in (5.46) is to be interpreted as a Strantonovich integral. Then the normal rules of calculus apply, and $Y = f(X)$ satisfies

$$dY = f'(X) b(X, t)\, dt + f'(X)\sigma(X, t) \circ dB.$$

The derivation of the Fokker-Planck equation is not as simple as for the Itô SDE, since the expected value of a Strantonovich integral is, in general, nonzero, but one can show that the Fokker-Planck equation for (5.63) is

$$\begin{aligned} \frac{\partial p}{\partial t} &= -\frac{\partial}{\partial x}\left(b\left(x, t\right) p\right) + \frac{1}{2} \frac{\partial}{\partial x}\left[\sigma(x, t) \frac{\partial}{\partial x}\left(\sigma\left(x, t\right) p\right)\right] \\ &= -\frac{\partial}{\partial x}\left\{\left[b\left(x, t\right) + \frac{1}{2}\sigma\left(x, t\right) \frac{\partial \sigma}{\partial x}\left(x, t\right)\right] p\right\} + \frac{1}{2} \frac{\partial^2}{\partial^2 x}\left[\sigma^2(x, t) p\right]. \end{aligned}$$

If σ is not constant, this PDE has a different drift term than the one in (5.60) arising from the Itô SDE (5.54).

Equivalently, the solution $X(t)$ of the Strantonovich SDE (5.63) is the same as the solution of the Itô SDE

$$dX = \left[b(X,t) + \frac{1}{2}\sigma(X,t)\frac{\partial \sigma}{\partial X}(X,t) \right] dt + \sigma(X)\, dB$$

with a corrected drift. Thus, the difference in drifts is simply a consequence of the difference in the definitions of the Itô and Strantonovich integrals, and it has no other significance. Of course, in using an SDE to model a system, one must choose an appropriate drift and noise, The drift will therefore depend on what definition of the stochastic integral one uses (see Remark 5.21).

8. Financial models

In this section we describe a basic SDE models of a financial market and derive the Black-Scholes formula for options pricing.

8.1. Stock prices

A simple model for the dynamics of the price $S(t) > 0$ of a stock at time t, introduced by Samuelson (1965), is provided by the Itô SDE

$$dS = \mu S\, dt + \sigma S\, dB \tag{5.64}$$

where μ and σ are constant parameters.

The drift-constant μ in (5.64) is the expected rate of return of the stock; in the absence of noise, $S(t) = S_0 e^{\mu t}$. The noise term describes random fluctuations in the stock price due to the actions of many individual investors. The strength of the noise is σS since we expect that the fluctuations in the price of a stock should be proportional to its price. The diffusion-constant σ is called the volatility of the stock; it is larger for more speculative stocks. Typical values for σ are in the range 0.2–0.4 in units of $(\text{years})^{1/2}$, corresponding to a standard deviation in the relative stock price of 20–40 percent per annum.

The dependence of the noise in (5.64) on the solution S differs from the constant noise in the Ornstein-Uhlenbeck SDE, which describes physical systems in thermodynamic equilibrium where the noise is fixed by the temperature.

As can be verified by the use of Itô's formula, the exact solution of (5.64) is

$$S(t) = S_0 \exp\left[\left(\mu - \frac{1}{2}\sigma^2 \right) t + \sigma B(t) \right] \tag{5.65}$$

where S_0 is the initial value of S. The process (5.65) is called *geometric Brownian motion*. The logarithm of $S(t)$ is Gaussian, meaning that $S(t)$ is lognormal. From (5.65) and (5.59), the expected value of $S(t)$ is

$$\mathbf{E}\left[S(t)\right] = \mathbf{E}\left[S_0\right] e^{\mu t},$$

consistent with what one obtains by averaging (5.64) directly.

Remark 5.21. We could equally well model the stock price by use of a Strantonovich SDE with a corrected value for the drift

$$dS = \left(\mu - \frac{1}{2}\sigma^2 \right) S\, dt + \sigma S\, \partial B. \tag{5.66}$$

The growth rate of the drift term in this equation is lower than the growth rate of the drift term in the corresponding Itô equation. This is because the Strantonovich noise contributes to the mean growth rate. Favorable fluctuations in the stock price increase the growth rate due to noise, and this outweighs the effect of unfavorable fluctuations that decrease the growth rate. The noise term in the Itô equation is defined so that its mean effect is zero. The solution of (5.66), which is found by the usual rules of calculus, is the same as the solution (5.65) of the corresponding Itô equation.

8.2. An ideal market

Consider, as an idealized model, a financial market that consists of single stock whose price $S(t)$ satisfies (5.64), and a risk-free security, such as a bond, whose price $R(t)$ satisfies the deterministic equation

$$dR = rR\,dt. \tag{5.67}$$

Thus, the value of the risk-free security is unaffected by random fluctuations in the stock market, and is assumed to have a fixed constant rate of return r.

We will refer to any item that is traded on the market, such as the stock, the bond, or a derivative, as a security. The prices, or values, of securities and the amounts owned by different traders are stochastic processes that are adapted to the filtration $\{\mathcal{F}_t : t \geq 0\}$ generated by the Brownian motion $B(t)$ in (5.64). This means that we cannot look into the future.

We assume that all processes have continuous sample paths. We further assume, for simplicity, that we can trade continuously without cost or restriction, that stocks and bonds are infinitely divisible, and that we can neglect any other complicating factors, such as dividends.

A portfolio is a collection of investments. If a portfolio consists of $a_i(t)$ units of securities with values $V_i(t)$, where $1 \leq i \leq n$, the value $\Pi(t)$ of the portfolio is

$$\Pi = \sum_{i=1}^{n} a_i V_i. \tag{5.68}$$

The value of the portfolio satisfies an SDE of the form

$$d\Pi = b\,dt + c\,dB.$$

We say that the portfolio is *risk-free* if $c = 0$, meaning that its value is not directly affected by random fluctuations in the market. Without further assumptions, however, the growth rate b could depend on B.

We say that the portfolio is *self-financing* if

$$d\Pi = \sum_{i=1}^{n} a_i\,dV_i. \tag{5.69}$$

As usual, this equation stands for the corresponding Itô integral equation. The condition (5.69) means that the change in the value of the portfolio is entirely due to the change in value of the securities it contains. Therefore, after the initial investment, no money flows in or out of the portfolio.

We will take as a basic assumption that the market allows no arbitrage opportunities in which traders can make a guaranteed profit through multiple transactions. Specifically, we assume that the value $\Pi(t)$ of any self-financing, risk-free security

must satisfy the ODE

$$d\Pi = r\Pi\, dt \tag{5.70}$$

where r is the risk-free rate of return in (5.67).

If there were a self-financing, risk-free portfolio whose instantaneous rate of return was higher (or lower) than the prevailing rate r, then traders could make a guaranteed profit by continuously buying (or selling) the securities in the portfolio. This would rapidly drive the rate of return of the portfolio and the prevailing rate r to the same value, which is the theoretical justification of the no-arbitrage assumption.

8.3. Derivatives

> It is a recipe for disaster to give one or two people complete authority to trade derivatives without a close monitoring of the risks being taken.[8]

Next, let us use this model to study the pricing of derivatives such as stock options. A derivative is a financial instrument that derives its value from some underlying asset. The asset could be almost anything, from pork bellies to next season's snowfall at a ski resort. Here, we consider derivatives that are contingent on the price of a stock.

We assume that the value $V(t)$ of the derivative is a deterministic function of the stock price $S(t)$ and the time t,

$$V(t) = f\left(S(t), t\right), \qquad f : (0,\infty) \times [0,\infty) \to \mathbb{R}.$$

Our aim is to determine what functions $f(S,t)$ provide values for a derivative that are consistent with the no-arbitrage assumption. The idea, following Black-Scholes (1973) and Merton (1973), is to construct a risk-free portfolio whose value replicates the value of the derivative.

Suppose that we sell, or write, one derivative, and form a portfolio that consists of:

(1) the derivative (whose value is a liability to us);
(2) a quantity $a(t)$ of the risk-free security with price $R(t)$;
(3) a quantity $b(t)$ of stock with price $S(t)$.

The value $\Pi(t)$ of the portfolio is given by

$$\Pi = aR + bS - V, \tag{5.71}$$

where R satisfies (5.67) and S satisfies (5.64).

We will choose $a(t)$, $b(t)$ so that the portfolio is self-financing and risk-free. In that case, its value must grow at the risk-free rate of return, and this will tell us how the value of the derivative $V(t)$ changes in terms of the price $S(t)$ of the stock.

The role of $R(t)$ is simply to provide a source of funds within the portfolio which allows us to adjust the stock holding as the value of the derivative fluctuates in order to maintain a risk-free position (this is called 'hedging'); R does not appear in the final result. If we did not include R in the portfolio, we would need to move funds in and out of the portfolio to make it risk-free.

From (5.69) and (5.71), the portfolio is self-financing if

$$d\Pi = a\, dR + b\, dS - dV. \tag{5.72}$$

[8] J. C. Hull, *Options Futures and Other Derivatives*, 4th ed., 2000.

Writing $V(t) = f\,(S(t), t)$, then using Itô's formula and (5.64), we get

$$dV = \left(\frac{\partial f}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 f}{\partial S^2}\right) dt + \frac{\partial f}{\partial S}\, dS.$$

Using this result and (5.67) in (5.72), we find that

$$d\Pi = \left(raR - \frac{\partial f}{\partial t} - \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 f}{\partial S^2}\right) dt + \left(b - \frac{\partial f}{\partial S}\right) dS.$$

Hence, the portfolio is risk-free if

$$b = \frac{\partial f}{\partial S}, \tag{5.73}$$

in which case

$$d\Pi = \left(raR - \frac{\partial f}{\partial t} - \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 f}{\partial S^2}\right) dt.$$

The no-arbitrage assumption (5.70) then implies that

$$d\Pi = r\Pi\, dt.$$

Equating these expressions for $d\Pi$, using (5.71) and (5.73), and simplifying the result, we find that

$$\frac{\partial f}{\partial t} + rS\frac{\partial f}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 f}{\partial S^2} = rf. \tag{5.74}$$

This is a PDE for $f(S, t)$, called the *Black-Scholes PDE*.

It is interesting to note that the rate of return μ of the stock in (5.64) does not appear in (5.74). The equation involves only the volatility σ of the stock and the risk-free rate of return r.

Equation (5.74) is a backward diffusion equation. In principle, any solution provides a feasible value-function for a derivative that is contingent on the stock. In the next section, we use (5.74) to determine the pricing of an option. The value of an option is known at the time when it comes to maturity, and this provides a final condition for (5.74). Solving the PDE backward in time then determines the initial price of the option. As we will see, although (5.74) has variable coefficients, we can obtain explicit solutions by transforming it to a constant coefficient heat equation.

8.4. Options

An option gives the holder the right, but not the obligation, to buy or sell an underlying asset for a specified price by a specified date. Options are primarily used to hedge against future risks or, perhaps more frequently, as a means of speculation.

The first reported use of options[9] seems to be by Thales who, after predicting a large olive crop by astronomical or astrological means, purchased one winter the right to use all the olive presses in Miletus and Chios for the coming harvest. When the large crop materialized, Thales was able to resell his rights for much more than they had cost him. Later on, tulip bulb options were heavily traded in Holland during the tulip mania of 1636 (until the market collapsed in 1637). Options were first traded on an organized exchange in 1973, on the Chicago Board Options Exchange. Since then the trading of options has developed into a global financial market of enormous volume.

[9]Aristotle, *Politics* I xi, 332 B.C.

There are two main types of options: a call option gives the holder the right to buy the underlying asset, while a put option give the holder the right to sell the asset. In an American option this right can be exercised at any time up to the expiration date; in a European option, the right can be exercised only on the expiration date itself.

Any options contract has two parties. The party who buys the option, is said to take the long position, while the party who sells, or writes, the option is said to take the short position. The writer receives cash up front, but has potential liabilities later on if the holder exercises the option. The holder incurs an immediate cost, but has the potential for future gains.

Let us consider, as an example, a European call option which gives the holder the right to buy a unit of stock at a prearranged price $K > 0$, called the *strike price*, at a future time $T > 0$. In this case, the value, or payoff, of the option at the expiration time T for stock price S is

$$f(S,T) = \max\{S - K, 0\}. \tag{5.75}$$

If $S \leq K$, the option is worthless, and the holder lets it expire; if $S > K$, the holder exercises the option and makes a profit equal to the difference between the actual price of the stock at time T and the strike price. We want to compute the fair value of the option at an earlier time.

To do this, we solve the Black-Scholes PDE (5.74) for $t \leq T$ subject to the final condition (5.75). We can find the solution explicitly by transforming (5.74) into the heat equation.

The change of independent variables $(S,t) \mapsto (x,\tau)$ given by

$$S = Ke^x, \qquad t = T - \frac{1}{\sigma^2}\tau \tag{5.76}$$

transforms (5.74) into the constant-coefficient equation

$$\frac{\partial f}{\partial \tau} = \frac{1}{2}\frac{\partial^2 f}{\partial x^2} + q\frac{\partial f}{\partial x} - \left(q + \frac{1}{2}\right) f \tag{5.77}$$

where

$$q = \frac{r}{\sigma^2} - \frac{1}{2}. \tag{5.78}$$

Since $0 < S < \infty$, we have $-\infty < x < \infty$. We have also reversed the time-direction, so that the final time $t = T$ corresponds to the initial time $\tau = 0$. The change of dependent variable in (5.77)

$$f(x,\tau) = Ke^{-qx-(q+1)^2\tau/2}u(x,\tau) \tag{5.79}$$

gives the heat equation

$$\frac{\partial u}{\partial \tau} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2}. \tag{5.80}$$

Rewriting (5.75) in terms of the transformed variables, we get the initial condition

$$u(x,0) = \begin{cases} e^{(q+1)x} - e^{qx} & \text{if } x > 0, \\ 0 & \text{if } x \leq 0. \end{cases} \tag{5.81}$$

The Green's function representation of the solution of (5.80)–(5.81) is

$$u(x,\tau) = \frac{1}{\sqrt{2\pi\tau}} \int_0^\infty \exp\left[-\frac{(x-y)^2}{2\tau}\right] \left[e^{(q+1)y} - e^{qy}\right] dy.$$

This integral is straightforward to evaluate by completing the square. For example,

$$\begin{aligned}\int_0^\infty \exp\left[-\frac{(x-y)^2}{2\tau}\right] e^{qy}\,dy &= \exp\left[qx+\frac{1}{2}q^2\tau\right]\int_0^\infty \exp\left[\frac{(y-x-q\tau)^2}{2\tau}\right]\,dy\\ &= \sqrt{\tau}\exp\left[qx+\frac{1}{2}q^2\tau\right]\int_{-\infty}^{\left(\frac{x+q\tau}{\sqrt{\tau}}\right)} e^{-z^2/2}\,dz\\ &= \sqrt{2\pi\tau}\exp\left[qx+\frac{1}{2}q^2\tau\right]\Phi\left(\frac{x+q\tau}{\sqrt{\tau}}\right)\end{aligned}$$

where Φ is the distribution function of the standard Gaussian,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-z^2/2}\,dz.$$

The function Φ is given in terms of the error function erf by

$$\Phi(x) = \frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right], \qquad \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-z^2}\,dz.$$

It follows that

$$\begin{aligned}u(x,\tau) = \exp\left[(q+1)x+\frac{1}{2}(q+1)^2\tau\right]\Phi\left(\frac{x+(q+1)\tau}{\sqrt{\tau}}\right)\\ -\exp\left(qx+\frac{1}{2}q^2\tau\right)\Phi\left(\frac{x+q\tau}{\sqrt{\tau}}\right).\end{aligned}$$

Using this equation in (5.79), then using (5.76) and (5.78) to rewrite the result in terms of the original independent variables (S,t), we get

$$f(S,t) = S\Phi\left(a(S,t)\right) - Ke^{-r(T-t)}\Phi\left(b(S,t)\right) \tag{5.82}$$

where

$$\begin{aligned}a(S,t) &= \frac{1}{\sigma\sqrt{T-t}}\left[\log\left(\frac{S}{K}\right)+\left(r+\frac{1}{2}\sigma^2\right)(T-t)\right],\\ b(S,t) &= \frac{1}{\sigma\sqrt{T-t}}\left[\log\left(\frac{S}{K}\right)+\left(r-\frac{1}{2}\sigma^2\right)(T-t)\right].\end{aligned} \tag{5.83}$$

Equation (5.82)–(5.83) is the Black-Scholes formula for the value $f(S,t)$, at stock-price S and time t, of a European call option with strike price K and expiration time T. It also involves the risk-free rate of return r of the market and the volatility σ of the underlying stock.

Other types of options can be analyzed in a similar way. American options are, however, more difficult to analyze than European options since the time, if any, at which they are exercised is not known *a priori*.

Bibliography

[1] M. J. Ablowitz, and A. Zeppetella, Explicit solutions of Fisher's equation for a special wave speed, *Bulletin of Mathematical Biology* **41** (1979), 835–840.

[2] W. Amrein, A. Hinz, and D. Pearson, *Sturm-Liouville Theory, Past and Present*, Birkhäuser, 2005.

[3] S. Antman, The equations for large vibrations of strings, *The American Mathematical Monthly*, **87** 1980, 359–370.

[4] S. Antman, *Nonlinear Problems of Elasticity*, 2nd ed., Springer-Verlag, New York, 2005.

[5] D. Applebaum, Lévy processes — from probability to finance and quantum groups, *Notices of the Amer. Math. Soc.* **51** (2004), 1336-1347.

[6] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Second Edition, Springer-Verlag, New Yok, 1989.

[7] G. I. Barenblatt, *Scaling*, Cambridge University Press, Cambridge, 2003.

[8] G. Batchelor, *An introduction to Fluid Mechnics*, Cambridge University Press, Cambridge, 1967.

[9] G. Batchelor, *The Theory of Homogeneous Turbulence*, Cambridge University Press, Cambridge, 1953.

[10] P. W. Bridgeman, *Dimensional Analysis*, Yale University Press, 1922.

[11] G. Buttazzo, M. Giaquinta and S. Hildebrandt, *One-dimensional Variational Problems*, Oxford Science Publications, 1998.

[12] G. Buttazzo and B. Kawohl, On Newton's problem of minimal resistances, *Mathematical Intelligencier* **15** No. 4 (1993), 7–12.

[13] B. Dacorogna, *Introduction to the Calculus of Varitions*, Imperial College Press, 2004.

[14] J. Dieudonné, *Foundations of Mathematical Analysis*, Vol.1, Academic Press, 1969.

[15] R. Durrett, *Stochastic Calculus: A Practical Introduction*, CRC Press, 1996.

[16] M. S. P. Eastham, *The Spectral Theory of Periodic Differential Equations*, Scottish Academic Press, 1973.

[17] P. Embrechts and M. Maejima, *Selfsimilar Processes* Princeton University Press, 2002.

[18] L. C. Evans, *Partial Differential Equations*, AMS, 1998.

[19] L. C. Evans, *An Introduction to Stochastic Differential Equations*, available at: `http://math.berkeley.edu/ evans/SDE.course.pdf`.

[20] G. L. Eyink and K. R. Sreenivasan, Onsager and the theory of hydrodynamic turbulence, *Rev. Modern Phys.* **78** (2006), 87-135.

[21] R. Feynman, and A. Hibbs, *Quantum Mechanics and Path Integrals*, McGraw-Hill, 1965.

[22] R. A. Fisher, The wave of advance of advantageous genes, *Ann. Eugenics* **7** (1937), 353369.

[23] U. Frisch, *Turbulence: The Legacy of A. N. Kolmogorov*, Cambridge University Press, 1995.

[24] H. Goldstein, *Classical Mechanics*.

[25] M. E. Gurtin, *Introduction to Continuum Mechanics*, Academic Press, New York, 1981.

[26] D. D. Holm, *Geometric Mechanics*, Imperical College Press, 2008.

[27] L. Hörmander, *The Analysis of Linear Partial Differential Operators I*, Second Edition, Springer-Verlag, Berlin, 1990.

[28] J. D. Jackson, *Classical Electrodynamics*.

[29] M. Kac, Can one hear the shape of a drum?, *American Mathematical Monthly* **73** 1966, 1–23.

[30] J. Keizer, *Statistical Thermodynamics of Nonequilibrium Processes*, pringer-Verlag, 1987.

[31] C. Kittel, *Introduction to Solid State Physics*.

[32] P. Kloeden, and E. Platen, *Numerical Solution of Stochastic Differential Equations*.

[33] A. Kolmogorov, I. G. Petrovskii, and N. S. Piskunov, A study of the diffusion equation with increase in the amount of substance, and its application to a biological problem. In editor,

Selected Works of A. N. Kolmogorov, Vol. I, ed. V. M. Tikhomirov, 242–270, Kluwer, 1991 (translated from *Bull. Moscow Univ., Math. Mech.* **1**, (1937) 1–25).
[34] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, 2nd ed., Pergamon Press, 1987.
[35] N. N. Lebedev, *Special Functions and their Applications*, Dover, New York, 1972.
[36] J. F. Marko, and E. Siggia, Stretching DNA, *Macromolecules* **26** (1995), 8759–8770.
[37] J. Mawhin and M. Willem, *Critical Point Theory and Hamiltonian Systems*, Springer-Verlag, 1989.
[38] R. Newburgh, J. Peidle, and W. Rueckner, Einstein, Perrin, and the reality of atoms: 1905 revisited, *Am. J. Phys.*, **74** (2006), 478–481.
[39] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
[40] P. Olver, *Applications of Lie Groups to Differential Equations*, Second Edition, Springer-Verlag, New York, 1993.
[41] P. Olver, *Equivalence, Invariants, and Symmetry*, Cambridge University Press, New York, 1995.
[42] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, reprinted edition, Springer-Verlag, New York, 1984.
[43] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.
[44] R. S. Strichartz, *A Guide to Distribution Theory and Fourier Transforms*, World Scientific, 2003.
[45] S. R. S. Varadhan, *Probability theory*, Courant Lecture Notes, AMS, 2001.
[46] S. R. S. Varadhan, *Stochastic Process*, Courant Lecture Notes, AMS, 2007.
[47] G. B. Whitham, *Linear and Nonlinear Waves*, John Wiley & Sons, New York, 1974.
[48] D. Yong, Strings,chains and ropes, *SIAM Review*, **48** 2006, 771-781