Partial Differential Equations

William G. Faris

May 17, 1999

Contents

Chapter 1

The Laplace equation

1.1 Gradient and divergence

In these lectures we follow the notation suggested by Evans. If u is a scalar function, then Du is the gradient, the vector of partial derivatives. If $\mathbf{J}$ is a vector function, then $D\mathbf{J}$ is the matrix of partial derivatives of the components of $\mathbf{J}$. The trace of this matrix is $\operatorname{div}\mathbf{J} = \operatorname{tr}(D\mathbf{J})$, the divergence of the vector field. (Similarly, the determinant of this matrix is the Jacobian.)

The matrix of second derivatives of u is D^2u, the Hessian matrix. This is a symmetric matrix. The trace of the Hessian matrix is the Laplacian:

$$\Delta u = \operatorname{tr}(D^2u) = \operatorname{div} Du. \tag{1.1}$$

We recall some basic calculus ideas. Let $\mathbf{J}$ be a vector field with divergence $\operatorname{div}\mathbf{J}$. The divergence theorem says that for bounded open sets V with smooth boundary ∂V we have

$$\int_V \operatorname{div}\mathbf{J}\,dy = \int_{\partial V} \mathbf{J}\cdot\nu\,dS, \tag{1.2}$$

where ν is the unit normal vector pointing outward, and dS is the surface measure.

Let s be a scalar function with gradient Ds. Then we have the product rule for differentiating in the form

$$\operatorname{div}(s\mathbf{J}) = Ds\cdot\mathbf{J} + s\operatorname{div}\mathbf{J}. \tag{1.3}$$

If we apply the divergence theorem to the left hand side of this equation, we get the fundamental integration by parts result:

$$\int_V Ds\cdot\mathbf{J}\,dy + \int_V s\operatorname{div}\mathbf{J}\,dy = \int_{\partial V} s\mathbf{J}\cdot\nu\,dS. \tag{1.4}$$

This will be used over and over again.

1.2 Equilibrium conservation laws

We begin with the most classical of partial differential equations, the Laplace equation. This equation is linear of second order, and is both translation and rotation invariant. It describes equilibrium in space. We will see in this first chapter that even though the equation only involves second derivatives, every solution automatically has all derivatives.

Now we proceed to the derivation of the Poisson and Laplace equations. Let $\mathbf{J}$ be a vector field on an open subset U of $\mathbf{R}^n$. This is the current. Let f be a function on U. This is the source (the rate of production of some quantity). An equilibrium conservation law is an equation of the form

$$\int_{\partial V} \mathbf{J} \cdot \nu \, dS = \int_V f \, dx. \tag{1.5}$$

Here V is supposed to range over suitable bounded open subsets of U. The boundary ∂V of each V is supposed to be a smooth subset of U. This equation says that the amount of substance flowing out of the region V is equal to the rate of production.

If we apply the divergence theorem, we obtain

$$\int_V \operatorname{div} \mathbf{J} \, dx = \int_V f \, dx. \tag{1.6}$$

Since this is assumed to hold for all subregions V, we have the differential form of the equilibrium conservation law:

$$\operatorname{div} \mathbf{J} = f. \tag{1.7}$$

Now assume that the current is proportional to the negative of the gradient of some scalar function u defined on U. Thus

$$\mathbf{J} = -Du. \tag{1.8}$$

We get the Poisson equation

$$-\triangle u = -\operatorname{div} Du = f. \tag{1.9}$$

When there is equilibrium with no source, then this is the Laplace equation

$$\triangle u = 0. \tag{1.10}$$

Even if one is interested in the Poisson equation, the Laplace equation is important, since the difference of two solutions of the Poisson equation is a solution of the Laplace equation. In the following we will usually think of the Poisson or Laplace equation being satisfied for a function u that is C^2 on some open set U.

In applications u could be temperature, density, or electric potential. The corresponding current $\mathbf{J}$ would be heat flux, diffusive current, or electric field.

The source f could be a source of heat, a source of diffusing particles, or an electric charge density.

If the Laplace equation is satisfied in U, then we have

$$\int_{\partial V} \mathbf{J} \cdot \nu \, dS = 0 \tag{1.11}$$

for every suitable subregion V of U. This says that there is no net flow into or out of the region V.

1.3 Polar coordinates

It will sometime be convenient to calculate integrals in polar coordinates. Thus

$$\int_{\mathbf{R}^n} f(x) \, dx = \int_0^\infty \int_{\partial B(x_0,r)} f(x) \, dS dr. \tag{1.12}$$

Here dS represents surface measure on the $n-1$ dimensional sphere $\partial B(x_0, r)$ of radius r centered at x_0. The total surface measure of the sphere is proportional to r^{n-1} and the proportionality constant will be taken so that it is by definition $n\alpha(n)r^{n-1}$. Thus, for example, $n\alpha(n)$ in dimensions $n = 1, 2, 3$ has the values 2, 2π, and 4π. In dimensions $n = 1, 2, 3$ these numbers represent the count of two points, the length of a unit circle, and the area of a unit sphere.

As an example, we can take $f(x) = \exp(-x^2)$ and $x_0 = 0$. Then

$$\int \exp(-x^2) \, dx = n\alpha(n) \int_0^\infty \exp(-r^2) r^{n-1} \, dr. \tag{1.13}$$

Here the total surface measure of the ball is defined to be $n\alpha(n)r^{n-1}$. We can also write this as

$$\int \exp(-x^2) \, dx = n\alpha(n) \frac{1}{2} \int_0^\infty u^{\frac{n}{2}-1} \exp(-u) \, du = n\alpha(n) \frac{1}{2} \Gamma(\frac{n}{2}). \tag{1.14}$$

When $n = 2$ this says that the value of the integral is π. It follows by factoring the exponential that for arbitrary dimension the value of the integral is $\pi^{\frac{n}{2}}$. Thus

$$\pi^{\frac{n}{2}} = n\alpha(n) \frac{1}{2} \Gamma(\frac{n}{2})). \tag{1.15}$$

This proves the basic fact that the area of the unit $n-1$ sphere is

$$n\alpha(n) = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} \tag{1.16}$$

The function $\Gamma(z)$ mentioned in this result is the usual Gamma function

$$\Gamma(z) = \int_0^\infty u^{z-1} e^{-u} \, du. \tag{1.17}$$

Its main properties are $\Gamma(z+1) = z\Gamma(z)$, $\Gamma(1) = 1$, and $\Gamma(1/2) = \sqrt{\pi}$.

We can also compute the volume of the unit ball by integrating the constant one over the ball in polar coordinates. The volume of the unit ball is thus

$$\alpha(n) = \frac{2\pi^{\frac{n}{2}}}{n\Gamma(\frac{n}{2})} = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}. \tag{1.18}$$

1.4 The mean value property

If $\alpha(n)$ is the n dimensional measure of the unit ball in $\mathbf{R}^n$, then $n\alpha(n)$ is the corresponding $n-1$ dimensional measure of the unit sphere in $\mathbf{R}^n$. Thus when $n = 1, 2, 3$ the values of $\alpha(n)$ are $2, \pi, (4/3)\pi$, while the values of $n\alpha(n)$ are $2, 2\pi, 4\pi$. The factor n comes of course from differentiating r^n to get nr^{n-1}.

In the following we write

$$\fint_{\partial B(x,r)} f(y)\, dS(y) = \frac{1}{n\alpha(n)r^{n-1}} \int_{\partial B(x,r)} f(y)\, dS(y) = \frac{1}{n\alpha(n)} \int_{\partial B(0,1)} f(x+rz)\, dS \tag{1.19}$$

for the mean value of f over the sphere of radius r. The last expression says that the mean value over the sphere is the integral of f over angles divided by the integral of 1 over angles. Also, we write

$$\fint_{B(x,r)} f(y)\, dy = \frac{1}{\alpha(n)r^n} \int_{B(x,r)} f(y)\, dy = \frac{1}{\alpha(n)} \int_{B(x,1)} f(x+rz)\, dz \tag{1.20}$$

for the mean value of f over a ball of radius r.

It will be very useful in the following to have the derivatives of these expressions with respect to the radius. They are obtained by differentiating under the integral sign.The fact that these are mean values is very important. The derivative of the average over the sphere is

$$\frac{d}{dr} \fint_{\partial B(x,r)} f(y)\, dS(y) = \fint_{\partial B(x,r)} Df(y) \cdot \frac{y-x}{r}\, dS(y). \tag{1.21}$$

The derivative of the average over the ball is

$$\frac{d}{dr} \fint_{B(x,r)} f(y)\, dy = \fint_{B(x,r)} Df(y) \cdot \frac{y-x}{r}\, dy. \tag{1.22}$$

Remark: The most useful formula for the derivative of the integral over the ball has a quite different nature. It says that

$$\frac{d}{dr} \int_{B(x,r)} f(y)\, dy = \int_{\partial B(x,r)} f(y)\, dS. \tag{1.23}$$

It is derived by differentiating in polar coordinates. Notice that in this form it does not involve a derivative of f. (Why not? What happens if one differentiates under the integral sign?)

Now we begin to apply this to the Poisson and Laplace equations.

Lemma 1.1 *Let $-\triangle u = f$. Fix x and let*

$$\phi(r) = \fint_{\partial B(x,r)} u(y)\, dS(y). \tag{1.24}$$

be the average of u over the sphere of radius r centered at x. Then

$$\phi'(r) = -\frac{r}{n} \fint_{B(x,r)} f(y)\, dy. \tag{1.25}$$

Proof: Use $Du = -\mathbf{J}$ and $\operatorname{div} \mathbf{J} = f$ to get

$$\phi'(r) = -\fint_{\partial B(x,r)} \mathbf{J}(y) \cdot \nu(y)\, dS(y) = -\frac{r}{n} \fint_{B(x,r)} f(y)\, dy. \tag{1.26}$$

Theorem 1.1 *Let u be a solution of the Laplace equation $\triangle u = 0$ in the open set U. Then for each closed ball $B(x,r) \subset U$ we have the mean value property:*

$$u(x) = \fint_{\partial B(x,r)} u(y)\, dS(y). \tag{1.27}$$

This says that the function is the average of its values over each sphere such that the ball is contained in the region.

Proof: From the lemma $\phi'(r) = 0$. So $\phi(r)$ is constant. However the limit of $\phi(r)$ as r tends to zero is $u(x)$. So $\phi(r) = u(x)$.

We could get the mean value property for balls as a corollary by doing the radial integral. However it is amusing to give an independent proof.

Lemma 1.2 *Let $-\triangle u = f$. Fix x and let*

$$\gamma(r) = \fint_{B(x,r)} u(y)\, dy. \tag{1.28}$$

be the average of u over the ball of radius r centered at x. Then

$$\gamma'(r) = -\frac{1}{2r} \fint_{B(x,r)} f(y)(r^2 - |y-x|^2)\, dy. \tag{1.29}$$

Proof: Use $Du = -\mathbf{J}$ and $\operatorname{div} \mathbf{J} = f$ to get

$$\gamma'(r) = -\frac{1}{r} \fint_{B(x,r)} \mathbf{J}(y) \cdot (y-x)\, dy = \frac{1}{2r} \fint_{B(x,r)} \mathbf{J} \cdot D(r^2 - |y-x|^2)\, dy. \tag{1.30}$$

The function $r^2 - |y-x|^2$ vanishes on the boundary, so we can integrate by parts and get the result.

Theorem 1.2 *Let u be a solution of the Laplace equation $\triangle u = 0$ in U. Then for each closed ball $B(x,r) \subset U$ we have the mean value property:*

$$u(x) = \fint_{B(x,r)} u(y)\, dy. \tag{1.31}$$

This says that the function is the average of its values over each ball contained in the region.

Theorem 1.3 *If u is a C^2 function that satisfies the mean value property for balls, then u is a solution of the Laplace equation.*

Proof: Let $-\triangle u = f$. Suppose that $f(x) \neq 0$ for some x. Let $\gamma(r)$ be the average over the ball of radius r about x. Then $\gamma'(r) = 0$. On the other hand, for r sufficiently small and for all y in $B(x, r)$ the value $f(y) \neq 0$ has the same sign as $f(x)$. The lemma then implies that $\gamma'(r) \neq 0$. This is a contradiction. Thus $f = 0$.

It is easy to see that the mean value property for spheres implies the mean value property for balls. Just to complete the circle, let us show that the mean value property for balls implies the mean value property for spheres.

Theorem 1.4 *The mean value property for balls implies the mean value property for spheres.*

Proof: Compute the derivative of the mean value over the ball by using the product rule:

$$0 = \frac{d}{dr}\frac{1}{\alpha(n)r^n}\int_{B(x,r)} u\,dy = -\frac{n}{\alpha(n)r^{n+1}}\int_{B(x,r)} u\,dy + \frac{1}{\alpha(n)r^n}\int_{\partial B(x,r)} u\,dS. \tag{1.32}$$

This equation may be solved for the mean value over the sphere.

1.5 Approximate delta functions

The delta function itself, of course, is not a function, but instead defines a measure $\delta(x)\,dx$ that assigns the value

$$\int f(y)\delta(y)\,dy = f(0) \tag{1.33}$$

to each continuous function f.

Let $\delta_1(x) \geq 0$ be a function on $\mathbf{R}^n$ with total integral one. Let $\delta_\epsilon(x) = \delta_1(x/\epsilon)/\epsilon^n$. Then this family of functions for $\epsilon > 0$ will be called an approximate delta function.

Recall that the convolution of two integrable functions f and g is an integrable function $f * g$ given by

$$(f * g)(x) = \int f(x-y)g(y)\,dy = \int g(x-y)\,f(y)\,dy. \tag{1.34}$$

The convolution product is commutative.

The following theorem justifies the terminology of approximate delta function.

Theorem 1.5 *Let f be a bounded continuous function. Then for each x the limit of the convolutions*

$$(\delta_\epsilon * f)(x) = \int \delta_\epsilon(x-y)f(y)\,dy \tag{1.35}$$

is

$$\lim_{\epsilon \to 0}(\delta_\epsilon * f)(x) = f(x). \tag{1.36}$$

Proof: Since the total integral of the approximate delta function is one, we have the identity

$$(\delta_\epsilon * f)(x) - f(x) = \int [f(x-y) - f(x)]\delta_\epsilon(y)\,dy \tag{1.37}$$

Fix x. Let $\lambda > 0$ be an arbitrary small positive number. Choose $\rho > 0$ so small that $|f(x-y) - f(x)| < \lambda/2$ for $|y| < \rho$. Then choose $\epsilon > 0$ so small that

$$\int_{|y| \geq \rho} \delta_\epsilon(y)\,dy = \int_{|z| \geq \rho/\epsilon} \delta_1(z)\,dz < \lambda/(4M), \tag{1.38}$$

where M is the bound on the absolute value of f. It follows that for all such $\epsilon > 0$ we can write the integral as a sum over $|y| < \rho$ and $|y| \geq \rho$ and get

$$|(\delta_\epsilon * f)(x) - f(x)| \leq \int |f(x-y) - f(x)|\delta_\epsilon(y)\,dy \leq \lambda/2 + \lambda/2 = \lambda. \tag{1.39}$$

Here are the standard examples of approximate delta functions. The most common is probably the Gaussian on $\mathbf{R}^n$:

$$\delta_\epsilon(x) = (2\pi\epsilon^2)^{-n/2}\exp(-\frac{x^2}{2\epsilon^2}). \tag{1.40}$$

We shall see that this plays a fundamental role in the solution of the heat equation in n space and one time dimensions. The relation is $\epsilon^2 = \sigma^2 t$ where σ^2 is the diffusion constant.

Another standard example is the Poisson kernel defined for x in $\mathbf{R}^{n-1}$. This is

$$\delta_\epsilon(x) = \frac{2\epsilon}{(n)\alpha(n)}\frac{1}{(x^2+\epsilon^2)^{n/2}}. \tag{1.41}$$

This gives the solution of the Laplace equation in a half space in n space dimensions. The relation is that ϵ is the space coordinate in the nth dimension.

An approximate delta function that is smooth with compact support is called a mollifier. Sometimes it is also very useful to require that it is a radial function. An example of a function with all these properties is the function $\delta_1(x)$ that is equal to

$$\delta_1(x) = C\exp(\frac{1}{1-r^2}). \tag{1.42}$$

within the unit ball and to zero outside it. Here the constant C is chosen so that the integral is one.

Theorem 1.6 *Suppose that f is an integrable function and that $\delta_\epsilon(x)$ is such a mollifier. Then the convolution $\delta_\epsilon * f$ is smooth. Furthermore, the value $\delta_\epsilon * f(x)$ at each x depends only on the values of f at a distance at most epsilon from x.*

The proof of the theorem requires justification of differentiation under the integral sign; however this is a consequence of the dominated convergence theorem.

1.6 Regularity

We now look at consequences of the mean value property. We will see that the consequences are both local and global. The local result is the regularity theorem that says that every solution of the Laplace equation is smooth. The global result is Liouville's theorem that every bounded solution of the Laplace equation on all of $\mathbf{R}^n$ is constant.

Theorem 1.7 *Let U be an open set and let u be a continuous function on U that satisfies the mean value property on every sphere $\partial B(x,r) \subset U$. Then u is smooth, that is, has derivatives of all orders.*

Proof: Consider an approximate delta function $\delta_\epsilon(x)$ that is a radial mollifier. Thus it a radial function that is smooth and vanishes outside of the ϵ ball.

Write

$$u_\epsilon(x) = \int_U \delta_\epsilon(x-y)u(y)\,dy \tag{1.43}$$

for x in the set U_ϵ of points in U with distance greater than ϵ to the boundary ∂U. We can differentiate under the integral sign, so u_ϵ is smooth on U_ϵ.

We can also write the convolution as

$$u_\epsilon(x) = \int_{B(0,\epsilon)} u(x-y)\delta_\epsilon(y)\,dy. \tag{1.44}$$

Now since the function δ_ϵ is radial, one can average over each sphere and then do the radial integral. The result is that

$$u_\epsilon(x) = \int_0^\epsilon \delta_\epsilon(y) \int_{\partial B(0,r)} u(x-y)\,dS(y)\,dr = u(x) \int_{B(0,\epsilon)} \delta_\epsilon(y)\,dy = u(x). \tag{1.45}$$

The fact that $u(x) = u_\epsilon(x)$ is a variation of the mean value theorem for balls, in that it says that $u(x)$ is an average that is rotationally symmetric about x but varies with the radius in a more complicated way.

The conclusion is that $u = u_\epsilon$ on U_ϵ. Therefore u is smooth on U_ϵ. Since $\epsilon > 0$ is arbitrary, it follows that u is smooth on U.

1.7 Liouville's theorem

Theorem 1.8 *Every bounded solution of the Laplace equation on $\mathbf{R}^n$ is constant.*

Proof: Let u be a solution of Laplace's equation in $\mathbf{R}^n$ that is bounded above and below by a constant.

Let x and y be two points. Let the distance between them be denoted by a. Let r be some arbitrary number. Then $u(x)$ is the mean value of u over $B(x,r)$ and $u(y)$ is the mean value of u over $B(y,r)$. These regions have a large overlap, and the only contribution to $u(x) - u(y)$ is from the integral of u over $B(x,r) \setminus B(y,r)$ and from the integral of $-u$ over $B(y,r) \setminus B(x,r)$. Now $B(y,r) \setminus B(x,r)$ is a subset of $B(y,r) \setminus B(y,r-a)$. Its relative proportion of volume is bounded by $(r^n - (r-a)^n)/r^n \leq na/r$. The same estimate holds for $B(y,r) \setminus B(x,r)$. Thus $|u(x) - u(y)|$ is bounded by $2Cna/r$. Since r is arbitrary we must have $u(x) = u(y)$. Since the points are arbitrary, u must be constant.

This theorem has a corollary that has implications for the Poisson equation.

Theorem 1.9 *Let u and v be two solutions of the Poisson equation in $\mathbf{R}^n$ that each vanish at infinity. Then $u = v$.*

Proof: Each solution is continuous and zero at infinity, so it is also bounded. Therefore $u - v$ is bounded, and hence $u - v = C$. However $u - v$ vanishes at infinity, it follows that $C = 0$.

1.8 The maximum principle

In this section U is a bounded open set. The boundary of U is ∂U, and the closure of U is $\bar{U}$.

In this section u will be a function that is continuous on $\bar{U}$. Therefore it assumes its maximum value M at some point in $\bar{U}$. We shall see that if u is a solution of the Laplace equation in U, then it actually must assume its maximum value M at some point in ∂U.

Theorem 1.10 *Let U be a bounded open set. Let u be a continuous function on $\bar{U}$ that is a solution of the Laplace equation in U. Let M be the maximum value of u on $\bar{U}$. Then there is a point x in ∂U with $u(x) = M$.*

Proof: If for every point x in U we have $u(x) < M$, then there is nothing to prove. Otherwise there is a point x in U with $u(x) = M$. Since U is open, there is a ball $B(x,r)$ that is a subset of U. Then $u(x)$ is the average of the values of $u(y)$ over y in $B(x,r)$. Since each $u(y) \leq M$ and the average $u(x) = M$, it follows that each $u(y) = M$. This shows that $u(y) = M$ for all y in the ball. Now take r large enough so that the ball becomes arbitrarily close to some point y on the boundary. By continuity, $u(y) = M$.

The maximum principle is intuitive in terms of equilibrium heat flow. If the temperature were hottest in the interior, then there would have to be a net outward flow from this point, which contradicts the conservation law.

There is also a corresponding minimum principle. The maximum and minimum principle together give the following result.

Corollary 1.1 *Let U be a bounded open set and u be continuous on $\bar{U}$. Let $\triangle u = 0$ in U with $u = 0$ on ∂U. Then $u = 0$.*

The importance of this is that it gives a uniqueness result for the Poisson equation in bounded open sets.

Corollary 1.2 *Let U be a bounded open set and u be continuous on $\bar{U}$. Let $-\triangle u = f$ in U with $u = g$ on ∂U. Then u is uniquely determined by f and g.*

The maximum principle says that the maximum is assumed on the boundary, but it does not rule out the possibility that the maximum is also assumed at some interior point. However the strong maximum principle shows that this is a degenerate case. Recall that an open set U is connected if it is not the disjoint union of two non-empty open sets V and W. The following result is the strong maximum principle.

Theorem 1.11 *Let U be a bounded open set that is connected. Let u be a continuous function on $\bar{U}$ that is a solution of the Laplace equation in U. Let M be the maximum value of u. If there is a point x in U with $u(x) = M$, then u is constant in U.*

Proof: Let V be the set of all x in U such that $u(x) = M$. Then V is an open set. The reason is that if x is in V, then there is a ball $B(x, r)$ that is a subset of U. Then $u(x)$ is the average of the values of $u(y)$ over y in $B(x, r)$. Since each $u(y) \leq M$ and the average $u(x) = M$, it follows that each $u(y) = M$. This shows that the ball $B(x, r)$ is a subset of V.

Let W be the set of all x in U such that $u(x) < M$. Then W is also an open set. This is because u is a continuous function.

Suppose as in the statement of the theorem that V is not empty. Then since U is connected, W must be empty. So $u = M$ in all of U.

1.9 Differentiating an integral

When can we differentiate under the integral sign. The following theorem gives an answer. The result is presented for ordinary derivatives, but the technique has an obvious extension for partial derivatives.

Theorem 1.12 *Let $f(t, y)$ be integrable in y for each t, and assume that there is a function $g(t, y)$ that is integrable in y such that*

$$|\frac{\partial f}{\partial t}(t + h, y)| \leq g(t, y). \tag{1.46}$$

Then

$$\frac{d}{dt} \int f(t, y)\, dy = \int \frac{\partial f}{\partial t}(t, y)\, dy. \tag{1.47}$$

This theorem says roughly that if you can control the derivative, then it is permissible to differentiate under the integral sign.

Proof: Compute the difference quotient and use the mean value theorem

$$\int \frac{f(t+h,y)-f(t,y)}{h}\,dy = \int \frac{\partial f}{\partial t}(t+h^*,y)\,dy. \tag{1.48}$$

Here h^* is between 0 and h. We want to take h to zero and argue that the limit of the integral on the right is the integral of the limiting function. Under the hypothesis of the theorem, this follows from the dominated convergence theorem.

Example: Take the special case of three dimensions, and let $\phi(x) = 1/(4\pi|x|)$. The Laplacian of this is zero at every point except the one singular point, the origin. Does this one point make a difference? Yes! In the next section we shall be interested in taking the Laplacian of the equation

$$u(x) = \int \phi(x-y)f(y)\,dy. \tag{1.49}$$

It is not permissible to differentiate under the integral sign and conclude that this is zero. This is because there is no way to dominate the difference quotients that come up in the course of taking the second partial derivatives of the singular expression $\phi(x-y)$ in the integrand.

Example: Write instead the equivalent expression

$$u(x) = \int \phi(y)f(x-y)\,dy. \tag{1.50}$$

Take f to be a smooth function with compact support. Can we differentiate under the integral sign? Certainly yes. Even though the function $\phi(y)$ has a singularity at zero, it is integrable near zero, as may be seen by going to polar coordinates. So the theorem applies, and we get

$$\triangle u(x) = \int \phi(y)\triangle f(x-y)\,dy. \tag{1.51}$$

1.10 The fundamental solution

Next we shall see how to solve the Poisson equation in $\mathbf{R}^n$ explicitly. We define the fundamental solution of the Poisson equation to be a radial solution ϕ defined everywhere but at the origin. Furthermore, we require that the gradient flux $\mathbf{J} = -D\phi$ have total integral one over every sphere centered at the origin. It is easy to work this out. The Laplace operator in polar coordinates is

$$\triangle = \frac{1}{r^{n-1}}\frac{\partial}{\partial r}r^{n-1}\frac{\partial}{\partial r} + \frac{1}{r^2}L, \tag{1.52}$$

where L is the angular part. Thus the fundamental solution (for $n \neq 2$) is of the form $C/r^{n-2} + D$. The flux in the radial direction is thus the negative of the radial derivative, which is $(n-2)C/r^{n-1}$. For the integral condition to be satisfied, the flux must be

$$\mathbf{J}(x) = -D\phi(x) = \frac{1}{n\alpha(n)r^{n-1}} \frac{x}{r}. \tag{1.53}$$

This defines the fundamental solution up to an additive constant. When $n > 2$ it is customary to choose the constant so that the solution approaches zero at infinity. This gives the result

$$\phi(x) = \frac{1}{n\alpha(n)} \frac{1}{n-2} \frac{1}{r^{n-2}} \tag{1.54}$$

for $n \neq 2$. When $n = 2$ it is customary to take the fundamental solution to be

$$\phi(x) = -\frac{1}{2\pi} \log(r). \tag{1.55}$$

However this is somewhat arbitrary in applications, since the arbitrary constant depends on the units in which r is measured. When $n = 1$ a conventional choice is

$$\phi(x) = -\frac{1}{2}|x|. \tag{1.56}$$

However now the constant has been chosen so that the value at the origin is zero. When the dimension is two or less the fundamental solution is unbounded below.

The physical interpretation of the fundamental solution is that there is a point source at the origin given by a delta function. The resulting equilibrium flow outward is the same in all directions. In dimensions one and two this equilibrium can be achieved only at the price of a potential that becomes more and more negative.

In working with these fundamental solutions, it is important to realize that, even though they are singular at the origin in dimensions two or more, they are integrable near the origin. This is convenient, for instance, in justifying a passage to the limit using the dominated convergence theorem.

For further purposes, we want to define an approximate fundamental solution ϕ_ϵ defined by the same formula, but with r replaced by $r_\epsilon = \sqrt{r^2 + \epsilon^2}$. Note that $dr_\epsilon/dr = r/r_\epsilon$. It follows that

$$-D\phi(x) = \frac{1}{n\alpha(n)} \frac{r}{r_\epsilon^n} \frac{x}{r}. \tag{1.57}$$

Furthermore, we can compute the divergence of this by multiplying its length by r^{n-1}, taking the r derivative, and dividing by r^{n-1}. The result is

$$-\triangle\phi_\epsilon(x) = \delta_\epsilon(x) = \frac{1}{\alpha(n)} \frac{\epsilon^2}{r_\epsilon^{n+2}}. \tag{1.58}$$

This is easily seen to be an approximate delta function. The divergence theorem shows that the integral of the function over a ball of radius a is the integral of the current over the sphere of radius a, which works out to be $a^n/(a^2+\epsilon^2)^{n/2}$. This approaches 1 as a tends to infinity.

Theorem 1.13 *Let f be a C^2 function with compact support. Let ϕ be the fundamental solution. Then the convolution*

$$u(x) = (\phi * f)(x) = \int \phi(x-y) f(y)\, dy \tag{1.59}$$

is a solution of the Poisson equation

$$-\triangle u = f. \tag{1.60}$$

Proof: The assumption on f is stronger than is really needed, but it is convenient for the following step. We compute

$$-\triangle u(x) = \int \phi(x-y)(-\triangle) f(y)\, dy. \tag{1.61}$$

Now we approximate. Integrate by parts twice to prove that

$$\int \phi_\epsilon(x-y)(-\triangle) f(y)\, dy = \int \delta_\epsilon(x-y) f(y)\, dy. \tag{1.62}$$

Now let ϵ approach zero. The left hand side converges by the dominated convergence theorem. The right hand side converges by the standard property of an approximate delta function. This immediately gives the Poisson equation.

It is important to note that if f is bounded and has compact support, then for $n > 2$ the solution $u(x)$ goes to zero at the same rate as the fundamental solution. So for dimension $n > 2$ we have a unique solution that approaches zero.

This solution also shows that for $n > 2$ we have the following positivity property: if $f \geq 0$, then the solution u given by the formula also satisfies $u \geq 0$. This is satisfying from the point of view of the interpretation where u is temperature or density; it says that a meaningful equilibrium is reached. The question of how to interpret the formula when $n \leq 2$ will be clarified when we consider the heat equation.

1.11 Energy

Energy methods are very powerful, but are also subtle. Here is an introduction. As usual we work on bounded open sets U. We assume that everything is sufficiently smooth so that the calculations make sense.

Consider functions w defined on $\bar{U}$ with $w = g$ on ∂U. This is a class of possible temperature distributions each with boundary value g. Let f be a specified function. The energy of a function w is defined to be

$$I(w) = \frac{1}{2} \int_U |Dw|^2\, dx - \int_U f\, w\, dx. \tag{1.63}$$

Thus the energy of w is large if the gradient Dw is large in some average sense. However it is small if w is concentrated near the source f.

As a first use of energy methods, we prove uniqueness of the solution of the Poisson equation. This is the same thing as showing that the only solution of the Laplace equation with zero boundary conditions is the zero solution.

Theorem 1.14 *Let U be a bounded open set with smooth boundary. Let u be continuous on $\bar{U}$. Let $\triangle u = 0$ in U with $u = 0$ on ∂U. Then $u = 0$.*

Proof: In this case

$$I(u) = \frac{1}{2} \int_U |Du|^2 \, dx = -\frac{1}{2} \int_U u \triangle u \, dx = 0. \tag{1.64}$$

Thus $Du = 0$, so u is constant. Clearly the constant must be zero.

It is interesting that this same uniqueness result can be proved by either the maximum principle or by the energy method. In general, the maximum principle is most useful in problems that have a probability flavor, and the energy method is more useful in problems that have a mechanical interpretation.

How can we use energy methods to prove existence? The following result gives a first hint.

Theorem 1.15 *Let u be a function at which the energy function I assumes its minimum. Then u is a solution of the Poisson equation $-\triangle u = f$ on U with boundary value $u = g$ on ∂U.*

Proof: Let v be a smooth function with compact support in U. Then

$$I(u + tv) = I(u) + t[\int_U Du \cdot Dv \, dx - \int_U f \, v \, dx] + t^2 \frac{1}{2} \int_U |Dv|^2 \, dx. \tag{1.65}$$

So the directional derivative of I along v is

$$\int_U Du \cdot Dv \, dx - \int_U f \, v \, dx = 0 \tag{1.66}$$

since u is a minimum. We can integrate by parts to write this as

$$\int_U -\triangle u \, v \, dx - \int_U f \, v \, dx = 0. \tag{1.67}$$

Since v is arbitrary, we must have $-\triangle u - f = 0$.

This result suggests a proof of existence of the solution of the Poisson equation. If one could prove from some general principle that the minimum has to be assumed, then this would accomplish the purpose. The problem, of course, is that one has to look at all functions w of finite energy, and it is necessary to define this space carefully in order to prove the minimization property. This can be done, but it requires some background in functional analysis.

Chapter 2

The heat equation

2.1 Conservation laws

We continue with the heat equation. The time dependence is not only of interest in its own right, but it also gives a new perspective on the equilibrium solutions that do not depend on time.

Now we proceed to the derivation of the heat equation. Let $\mathbf{J}$ be a vector field on an open subset U of $\mathbf{R}^n$. This is the current. Let f be a function on U. This is the source (the rate of production of some quantity). A conservation law is an equation of the form

$$\frac{d}{dt}\int_V u\,dx + \int_{\partial V} \mathbf{J}\cdot\nu\,dS = \int_V f\,dx. \tag{2.1}$$

Here V is supposed to range over suitable bounded open subsets of U. The boundary ∂V of each V is supposed to be a smooth subset of U. This equation says that the rate of increase of u in the region plus the amount of substance flowing out of the region V is equal to the rate of production.

If we differentiate under the integral sign and also apply the divergence theorem, we obtain

$$\int_V \frac{\partial u}{\partial t}\,dx + \int_V \operatorname{div}\mathbf{J}\,dx = \int_V f\,dx. \tag{2.2}$$

Since this is assumed to hold for all subregions V, we have the differential form of the equilibrium conservation law:

$$\frac{\partial u}{\partial t} + \operatorname{div}\mathbf{J} = f. \tag{2.3}$$

Now assume that the current is proportional to the negative of the gradient of some scalar function u defined on U. Thus

$$\mathbf{J} = -\frac{1}{2}\sigma^2 Du, \tag{2.4}$$

where the diffusion constant $\sigma^2 > 0$. We get the heat equation

$$\frac{\partial u}{\partial t} = \frac{1}{2}\sigma^2 \triangle u + f. \tag{2.5}$$

This equation has a physical parameter, the diffusion constant. It has dimensions of squared distance divided by time. The physical meaning of this is that diffusion from a point is a rather slow process: the distance travelled is proportional on the average to the square root of the time.

The interpretation of the quantity in the conservation law depends on the application. One can think of the law as a law of conservation of mass. In that case, u is interpreted as a density, and $\mathbf{J}$ is a flow of particles. The source term f is a rate of particle production. The fact that the $\mathbf{J}$ is proportional to $-Du$ is Fick's law of diffusion.

One can also think of the law as a law of conservation of thermal energy. In that case u is the temperature, and $\mathbf{J}$ is a heat flow. In that interpretation f represents heat production. The equation that says that $\mathbf{J}$ is a constant times $-Du$ is Fourier's law of heat condution.

2.2 The fundamental solution

The fundamental solution of the heat equation is

$$\phi(x, t) = (2\pi\sigma^2 t)^{-\frac{n}{2}} \exp(-\frac{x^2}{2\sigma^2 t}), \tag{2.6}$$

defined for each $t > 0$. Here x is a point in $\mathbf{R}^n$. As usual, in applications we often are most interested in the cases $n = 1, 2, 3$, but the nice thing is that the formula is independent of dimension. Sometimes we may want to think of $\phi(x, t)$ as defined for $t < 0$; in this case we take it to be zero. Sometimes we want to think of this solution as a function of x parametrized by time $t > 0$, and in this case we write it as $\phi_t(x)$.

This is one of the most fundamental formulas in all of mathematics, and so it should be carefully memorized. It is the famous Gaussian or normal density of probability theory. In the version given here, $\phi_t(x)$ has mean equal to the zero vector and variance equal to $\sigma^2 t$ times the identity matrix. The normalization is chosen so that the total integral of $\phi_t(x)$ with respect to x is one. The interpretation of the variance is that the total integral of $x^2\phi_t(x)$ with respect to x is $n\sigma^2 t$. The n comes from adding the effects of the n components. Thus the average value of the squared distance is proportional to the time.

The relation to the heat equation is based on the two following facts. First, $\phi(x, t)$ is a solution of the heat equation in the region consisting of all x in $\mathbf{R}^n$ and all $t > 0$. Second, the functions $\phi_t(x)$ form an approximate delta function as $t \to 0$. These two facts are enough to prove the following theorem.

Theorem 2.1 *Let $g(x)$ be bounded and continuous. The function*

$$u(x, t) = (\phi_t * g)(x) = \int \phi_t(x - y) g(y)\, dy \tag{2.7}$$

is a solution of the homogeneous heat equation for $t > 0$ satisfying the initial condition $u(x, 0) = g(x)$.

This solution tells how an initial distribution of heat spreads out in time. It is remarkable that the fundamental solution also gives the effect of a source of heat.

Theorem 2.2 *Let $f(x, t)$ be bounded and continuous for $t > 0$ (and zero for $t \leq 0$). Then*

$$u(x, t) = (\phi * f)(x, t) = \int_0^t \phi(x - y, t - s) f(y, s)\, dy\, ds \tag{2.8}$$

is a solution of the inhomogeneous heat equation with source f that also satisfies the initial condition $u(x, 0) = 0$.

This theorem says that the effect at later time t of a source acting at time s is that of an initial condition at time s spreading out over the time $t - s$.

Proof: We content ourselves with a formal calculation. For each s, let $v(x, t; s)$ be the solution of the homogeneous heat equation with initial condition $v(x, s; s) = f(x, s)$. Thus for $t > s$ we have

$$v(x, t; s) = \int \phi(x - y, t - s) f(y, s)\, dy. \tag{2.9}$$

Let

$$u(x, t) = \int_0^t v(x, t; s)\, ds. \tag{2.10}$$

If we differentiate according to the usual rules, we get

$$\frac{\partial}{\partial t} u(x, t) = \int_0^t \frac{\partial}{\partial t} v(x, t; s)\, ds + v(x, t; t) \tag{2.11}$$

This is the same as

$$\frac{\partial}{\partial t} u(x, t) = \frac{1}{2}\sigma^2 \triangle u(x, t) + f(x, t). \tag{2.12}$$

2.3 Approach to equilibrium

Say that the source is independent of time. Then the theorem takes the following form.

Theorem 2.3 *Let $f(x)$ be bounded and continuous. Then*

$$u(x, t) = (\phi * f)(x, t) = \int_0^t \phi(x - y, t') f(y)\, dy\, dt' \tag{2.13}$$

is a solution of the inhomogeneous heat equation with source f that also satisfies the initial condition $u(x, 0) = 0$.

The proof of this theorem is to make the change of variable $t' = t - s$ for each fixed t.

This theorem shows that it is of interest to study the time integrated fundamental solution. In order to compare it with previous results it is convenient to multiply by $\sigma^2/2$. Thus we consider

$$g_t(x) = \frac{\sigma^2}{2}\int_0^t \phi(x,t')\,dt'. \tag{2.14}$$

We can compute this explicitly by making the change of variables $a = x^2/(2\sigma^2 t)$. This gives

$$g_t(x) = \frac{1}{4\pi^{n/2}}\frac{1}{|x|^{n-2}}\int_{\frac{x^2}{2\sigma^2 t}}^{\infty} a^{\frac{n}{2}-2}e^{-a}\,da. \tag{2.15}$$

We can already see something interesting. When $n > 2$ this approaches the limit

$$g(x) = \frac{1}{4\pi^{n/2}}\Gamma(\frac{n}{2}-1)\frac{1}{|x|^{n-2}}. \tag{2.16}$$

This is the fundamental solution of the Laplace equation. Recall that the area of the unit sphere is

$$n\alpha(n) = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} = \frac{4\pi^{\frac{n}{2}}}{(n-2)\Gamma(\frac{n}{2}-1)}. \tag{2.17}$$

We have proved the following theorems.

Theorem 2.4 *When $n > 2$ the function $g_t(x)$ approaches the fundamental solution of the Laplace equation as $t \to \infty$.*

Theorem 2.5 *Consider dimension $n > 2$. Let $f(x)$ be a time-independent source, and let $u(x,t)$ be the solution of the corresponding heat equation $\partial u/\partial t = (\sigma^2/2)\triangle u + f$ with initial condition $u = 0$ at $t = 0$. Then the limit of $u(x,t)$ as $t \to \infty$ is the solution of the Poisson equation $(\sigma^2/2)\triangle u + f = 0$.*

Why does this result depend on dimension in this way? When $n \neq 2$ we can use integration by parts to write the integrated fundamental solution as as

$$g_t(x) = \frac{1}{2(n-2)\pi^{n/2}}\frac{1}{|x|^{n-2}}[\int_{\frac{x^2}{2\sigma^2 t}}^{\infty} a^{\frac{n}{2}-1}e^{-a}\,da - \left(\frac{x^2}{2\sigma^2 t}\right)^{\frac{n}{2}-1}\exp(-\frac{x^2}{2\sigma^2 t})]. \tag{2.18}$$

We can also write this as

$$g_t(x) = \frac{1}{2(n-2)\pi^{n/2}}[\frac{1}{|x|^{n-2}}\int_{\frac{x^2}{2\sigma^2 t}}^{\infty} a^{\frac{n}{2}-1}e^{-a}\,da - \left(\frac{1}{2\sigma^2 t}\right)^{\frac{n}{2}-1}\exp(-\frac{x^2}{2\sigma^2 t})]. \tag{2.19}$$

The integral now always has a finite limit as $t \to \infty$. Furthermore, when $n > 2$ the second term approaches zero as $t \to \infty$. However for $n \leq 2$ it is quite a different story.

Theorem 2.6 *When $n = 1$ the function*

$$g_t(x) = \frac{1}{2\pi^{1/2}}[-|x|\int_{\frac{x^2}{2\sigma^2 t}}^{\infty} a^{-\frac{1}{2}}e^{-a}\,da + (2\sigma^2 t)^{\frac{1}{2}}\exp(-\frac{x^2}{2\sigma^2 t})] \tag{2.20}$$

is the sum of a term that approaches $-(1/2)|x|$ with a term that approaches infinity at a rate proportional to $\sqrt{t}$ as $t \to \infty$.

What this shows is that there is no equilibrium for diffusion in one dimension. For large time, the solution $g_t(x)$ is approximately in the form of positive spike. The usual fundamental solution $-(1/2)|x|$ is only obtained at the price of subtracting off a positive infinite constant. What this says physically is that a steady source of heat never produces an equilibrium. Even though heat is being radiated away, there is a buildup of temperature at each point that never ceases.

There is a similar result for $n = 2$. This is obtained by the same kind of integration by parts. The form of the solution, however, is a bit different.

Theorem 2.7 *When $n = 2$ the function*

$$g_t(x) = \frac{1}{4\pi}[\int_{\frac{x^2}{2\sigma^2 t}}^{\infty} \log(a)e^{-a}\,da - \log\left(\frac{x^2}{2\sigma^2 t}\right)\exp(-\frac{x^2}{2\sigma^2 t})]. \tag{2.21}$$

is the sum of a term that approaches $-(1/2\pi)\log(|x|)$ with a term that approaches infinity at a rate proportional to $\log(t)$ as $t \to \infty$.

Thus there is also no equilibrium for diffusion in two dimensions. However this is very much a borderline case. Again a steady source of heat would produce a continually growing temperature at each point. However there is almost enough room for the heat to escape, so the rate of increase is quite slow.

In three dimensions, of course, there is plenty of room, and the heat simply escapes into the vast expanses of space.

2.4 The mean value property

For each x, t we define the heat ball $E(x, t, r)$ of radius r to be the set of all (y, s) with $s \leq t$ such that $\phi(x - y, t - s) \geq 1/r^n$. Notice that the heat ball around the point (x, t) is entirely in the past of the point.

How can one visualize this heat ball? Think of s as approaching t from below. When s is much less than t, then the variance $\sigma^2(t - s)$ is very large, and so there are no points in the ball at this early time. The first s for which the point $y = x$ is on the ball is when $2\pi\sigma^2(t - s) = r^2$. When s is close enough to t so that $2\pi\sigma^2(t - s) < r^2$, then the corresponding y such that (y, s) belong to the heat ball form an ordinary ball in space centered at x. As s continues to approach t these spatial balls grow to a maximum size and then begin go shrink. When $t - s$ is very small, then the spatial balls are also very small. Finally, when $s = t$, the spatial ball has shrunk to the point x again.

We want to write the equation for the boundary of the heat ball as the zero of a function. The natural function is

$$\psi(y,s) = \log(\phi(x-y,t-s)r^n). \tag{2.22}$$

In order to calculate with this function, we need to compute its partial derivatives. This is easy. We have

$$\frac{\partial\psi}{\partial s} = \frac{n}{2}\frac{1}{t-s} - \frac{(y-x)^2}{2\sigma^2(t-s)^2}. \tag{2.23}$$

Also

$$D\psi = -\frac{y-x}{\sigma^2(t-s)}. \tag{2.24}$$

From this we see that

$$-\frac{\partial\psi}{\partial s} + \frac{n}{2}\frac{1}{t-s} = \frac{1}{2(t-s)}(y-x)\cdot D\psi = \frac{(y-x)^2}{2\sigma^2(t-s)}. \tag{2.25}$$

This suggests the following notion of average over the unit ball.

Lemma 2.1 *The integral over each heat ball of* $1/(2\sigma^2 r^n)$ *times* $(x-y)^2/(t-s)^2$ *is one:*

$$M_r(1) = \frac{1}{r^n}\int\int_{E(x,t,r)} \frac{1}{2\sigma^2}\frac{(x-y)^2}{(t-s)^2}\,dy\,ds = 1. \tag{2.26}$$

Proof: The integral is over the region $\phi(x-y,t-s) \geq 1/r^n$. First make the change of variable $rz = x-y$ and $r^2\tau = t-s$. This gives the result

$$M_r(1) = \int\int_{\phi(z,\tau)\geq 1} \frac{z^2}{2\sigma^2\tau^2}\,dz\,d\tau. \tag{2.27}$$

Now change the relative scale of space and time by making the change of variable $w = z/\sqrt{\tau}$. The integral reduces to

$$M_r(1) = \int\int_{\phi(w,1)\geq\tau^{\frac{n}{2}}} \frac{w^2}{2\sigma^2}\tau^{\frac{n}{2}-1}\,dw\,d\tau. \tag{2.28}$$

Finally set $a = \tau^{\frac{n}{2}}$. This gives

$$M_r(1) = \int\int_{\phi(w,1)\geq a} \frac{w^2}{n\sigma^2}\,dw\,da = \int\int_0^{\phi(w,1)} da\frac{w^2}{n\sigma^2}\,dw = \int \phi(w,1)\frac{w^2}{n\sigma^2}\,dw = 1, \tag{2.29}$$

since the variance is σ^2 times the identity matrix.

Theorem 2.8 *If* u *is a solution of the homogeneous heat equation, then the value of* u *at each space time point* (x,t) *is the average of its values over each heat ball about this point. Thus*

$$u(x,t) = M_r(u) = \frac{1}{r^n}\int\int_{E(x,t,r)} u(y,s)\frac{1}{2\sigma^2}\frac{(x-y)^2}{(t-s)^2}\,dy\,ds. \tag{2.30}$$

Proof: The integral is over the region $\phi(x-y,t-s) \geq 1/r^n$. Make the change of variable $rz = x-y$ and $r^2\tau = t-s$. The integral becomes an integral over the unit heat ball:

$$M_r(u) = \int\int_{\phi(z,\tau)\geq 1} u(x-rz, t-r^2\tau)\frac{1}{2\sigma^2}\frac{z^2}{\tau^2}\,dz\,d\tau. \qquad (2.31)$$

Now differentiate with respect to r. We obtain

$$M_r'(u) = \int\int_{\phi(z,\tau)\geq 1} [Du(x-rz,t-r^2\tau)\cdot(-z) - \frac{\partial}{\partial s}u(x-rz,t-r^2\tau)2r\tau]\frac{1}{2\sigma^2}\frac{z^2}{\tau^2}\,dz\,d\tau. \qquad (2.32)$$

We can express this in the original coordinates as

$$M_r'(u) = \frac{1}{r^{n+1}}\int\int_{E(x,t,r)} [Du(y,s)\cdot(y-x) - \frac{\partial}{\partial s}u(y,s)2(t-s)]\frac{1}{2\sigma^2}\frac{(x-y)^2}{(t-s)^2}\,dy\,ds. \qquad (2.33)$$

Now we write this in terms of the function ψ that defines the heat ball. This becomes

$$M_r'(u) = \frac{1}{r^{n+1}}\int\int_{\psi>0} [-Du(y,s)\cdot(y-x)\frac{\partial\psi}{\partial s} - \sigma^2\frac{n}{2}Du(y,s)\cdot D\psi + \frac{\partial u(y,s)}{\partial s}(y-x)\cdot D\psi]\,dy\,ds. \qquad (2.34)$$

Now integrate by parts, and use the fact that ψ vanishes on the boundary. We obtain

$$M_r'(u) = \frac{1}{r^{n+1}}\int\int_{\psi>0} [\frac{\partial}{\partial s}Du(y,s)\cdot(y-x) + \sigma^2\frac{n}{2}\triangle u(y,s)\psi - \frac{\partial}{\partial s}Du(y,s)\cdot(y-x) - \frac{\partial u(y,s)}{\partial s}n]\psi\,dy\,ds. \qquad (2.35)$$

There is a cancellation, so we get

$$M_r'(u) = \frac{n}{r^{n+1}}\int\int_{\psi>0} [\frac{1}{2}\sigma^2\triangle u(y,s) - \frac{\partial u(y,s)}{\partial s}]\psi\,dy\,ds = 0. \qquad (2.36)$$

If follows that $M_r(u)$ is constant. By shrinking r to zero, we see that $M_r(u) = u(x,t)$.

2.5 The maximum principle

Let U be a bounded open set in $\mathbf{R}^n$. Let $T > 0$. Then the space-time cylinder U_T is defined to be $U \times (0,T]$. Thus it contains all points (x,t) in space-time such that x belongs to U and $0 < t \leq T$. The space-time boundary of the space-time cylinder is defined to be $\Gamma_T = \bar{U}_T \setminus U_T$. Thus it contains all points (x,t) in space such that either $t = 0$ and x is in U (initial points) or such that $0 \leq t \leq T$ and x is in ∂U (boundary points). Notice that the time T points in the region U do not belong to the space-time boundary.

Theorem 2.9 *Let u be a solution of the homogeneous heat equation in the space-time cylinder that is continous on the closure of the space-time cylinder. Let M be the maximum value of the solution on the closure of the space-time cylinder. Then there exists a point x on the space-time boundary with $u(x) = M$.*

Proof: If for every point (x, t) in the space-time interior $u(x, t) < M$, then there is nothing more to prove. So let (x, t) be a point in the space-time interior with $u(x, t) = M$. Then there is a heat ball $E(x, t, r)$ that is also in the space-time interior. Since $u(x, t)$ is the integral of $u(y, s)$ over the (y, s) in the heat ball, and each $u(y, s) \leq M$, it follows that each $u(y, s) = M$. Now r incrase until the heat ball $E(x, t, r)$ becomes arbitrarily close to the space-time boundary. By continuity, there is a point (y, s) on the space-time boundary with $u(y, s) = M$.

Note: There is also a strong maximum principle that can be proved under the additional assumption that U is connected.

Along with the maximum principle comes a minimum principle. This gives the following fundamental result.

Corollary 2.1 *Let U be a bounded open set and let $T > 0$. Let u be a solution of the homogeneous heat equation*

$$\frac{\partial u}{\partial t} = \frac{1}{2}\sigma^2 \triangle u \tag{2.37}$$

in U_T that is continuous on the closure and such that $u = 0$ on the space-time boundary Γ_T. Then $u = 0$.

Corollary 2.2 *Let U be a bounded open set and let $T > 0$. Consider the inhomogeneous heat equation*

$$\frac{\partial u}{\partial t} = \frac{1}{2}\sigma^2 \triangle u + f \tag{2.38}$$

in U_T with the boundary condition $u = g$ on the space-time boundary Γ_T. Then the solution u is uniquely determined by f and g.

This theorem says that the temperature is uniquely specified by the source f and by the boundary condition g given on $\bar{U}$ at $t = 0$ and on ∂U at $0 < t \leq T$.

Chapter 3

The wave equation

3.1 Mechanical conservation laws

Now we consider a system of conservation laws. The equation

$$\frac{d}{dt}\int_V u\,dx + \int_{\partial V} \mathbf{J}\cdot\nu\,dS = 0. \tag{3.1}$$

says that the rate of increase of the mass in a region is the amount of material that flows in. The differential form is

$$\frac{\partial u}{\partial t} + \operatorname{div}\mathbf{J} = 0. \tag{3.2}$$

This is a conservation law of the form that we have seen before.

However now the second conservation law is a vector equation. It says that

$$\frac{d}{dt}\int_V \mathbf{J}\,dx + \int_{\partial V} p\nu\,dS = \int_V \mathbf{F}\,dx. \tag{3.3}$$

This says that the acceleration is equal to the net inward pressure on the surface plus an extra force term. Notice that p is a scalar while ν, the outward normal, is a vector. The differential form of this law is

$$\frac{\partial \mathbf{J}}{\partial t} + Dp = \mathbf{F}. \tag{3.4}$$

It is a form of the equation of conservation of momentum. Notice that the gradient Dp comes from the divergence theorem! This can be shown by taking the dot product of both sides of the conservation law with a constant vector $\mathbf{a}$. Then one can apply the ordinary divergence theorem. The divergence that arises is $\operatorname{div}(p\mathbf{a}) = Dp\cdot\mathbf{a}$.

Finally, we need an equation that relates the pressure p to the density u, and this is

$$Dp = c^2 Du. \tag{3.5}$$

This says that a density gradient results in a pressure gradient.

These equations have as consequence the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \triangle u + f. \tag{3.6}$$

Here $f = -\operatorname{div} \mathbf{F}$.

In this interpretation the variable u is the density of a gas, and the current $\mathbf{J}$ is a fluid velocity. The wave equation of course has many other interpretations. For instance, one can take u to be an electric potential and $\mathbf{J}$ a magnetic potential. Then $-\mathbf{F}$ is the electric field, and $f = -\operatorname{div} \mathbf{F}$ is the charge density. In all cases c is the speed of propagation.

Since this is a second order equation, it seems reasonable to specify both the initial displacement $u = g$ at $t = 0$ and the initial velocity $\partial u/\partial t = h$ at $t = 0$.

From now on we shall concentrate on the task of solving the homogeneous wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \triangle u \tag{3.7}$$

with the initial conditions $u = 0$ and $\partial u/\partial t = h$ at $t = 0$. We shall refer to this solution as the fundamental solution. The next lemma shows that the time derivative of the fundamental solution gives the solution of the other initial value problem for the homogeneous equation. The lemma following that shows that the fundamental solution gives the solution of the inhomogeneous problem.

Lemma 3.1 *Let u be a solution of the homogeneous wave equation with $u = 0$ and $\partial u/\partial t = g$ at $t = 0$. Then $v = \partial u/\partial t$ is a solution of the homogeneous wave equation*

$$\frac{\partial^2 v}{\partial t^2} = c^2 \triangle v. \tag{3.8}$$

with the initial condition $v = g$ at $t = 0$ and $\partial v/\partial t = 0$ at $t = 0$.

Proof: It is obvious that the equation is satisfied and that $v = g$ at $t = 0$. On the other hand, $\partial v/\partial t = c^2 \triangle u$. Since $u = 0$ at $t = 0$, it follows that $\triangle u = 0$ at $t = 0$.

The inhomogeneous equation is taken care of by the following Duhamel formula.

Lemma 3.2 *Let $u(x, t; s)$ be a solution of the homogeneous wave equation for $t > s$ with initial conditions $u(x, s; s) = 0$ and $\partial u/\partial t(x, s; s) = f(x; s)$. Then*

$$v(x, t) = \int_0^t u(x, t; s)\, ds \tag{3.9}$$

is a solution of the inhomogeneous wave equation

$$\frac{\partial^2 v}{\partial t^2} = c^2 \triangle v + f. \tag{3.10}$$

with the initial conditions $v = 0$ at $t = 0$ and $\partial v/\partial t = 0$ at $t = 0$.

Proof: The first differentiation gives

$$\frac{\partial}{\partial t}v(x,t) = \int_0^t \frac{\partial}{\partial t}u(x,t;s)\,ds + u(x,t;t) \tag{3.11}$$

where $u(x,t;t) = 0$. The second differentiation gives

$$\frac{\partial^2}{\partial t^2}v(x,t) = \int_0^t \frac{\partial^2}{\partial t^2}u(x,t;s)\,ds + \frac{\partial u}{\partial t}(x,t;t). \tag{3.12}$$

where $\partial u/\partial t(x,t;t) = f(x,t)$.

3.2 The d'Alembert solution

We want to solve the one-dimensional wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}. \tag{3.13}$$

It is easy to see that every function of the form

$$u(x,t) = F(x+ct) + G(x-ct) \tag{3.14}$$

is a solution.

Let us compute the fundamental solution. If we want to have the initial condition $u(x,0) = 0$, then the solution must be of the form

$$u(x,t) = F(x+ct) - F(x-ct). \tag{3.15}$$

If we also want $\partial u/\partial t = h(x)$, then we get

$$2cF'(x) = h(x). \tag{3.16}$$

The solution is now easy; the result is the following d'Alembert formula for the solution.

Theorem 3.1 *The fundamental solution of the one-dimensional wave equation*

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2} \tag{3.17}$$

with $u = 0$ and $\partial u/\partial t = h$ at $t = 0$ is

$$u(x,t) = \frac{1}{2c}\int_{x-ct}^{x+ct} h(y)\,dy. \tag{3.18}$$

Notice that the solution is just t times the average over the ball of radius ct centered at x. Of course in one dimension this ball is an interval.

The solution of the equation with $u = g$ and $\partial u/\partial t = 0$ at $t = 0$ may be obtained by inserting g and differentiating with respect to t. The result is $u(x,t) = (1/2)[g(x+ct) + g(x-ct)]$. This is just the average over the sphere of radius ct centered at x. Of course in one dimension this sphere consists of two points.

We can also solve the inhomogeneous equation by the Duhamel formula

$$u(x,t) = \int_0^t \frac{1}{2c} \int_{x-c(t-s)}^{x+c(t-s)} f(y, t-s)\, dy\, ds. \tag{3.19}$$

This is the integral over a triangle in space-time with one vertex at the space-time point (x,t) and with the other two vertices at $(x \pm ct, 0)$. Thus the solution at (x,t) only depends on the source at (y,s), where the signal from (y,s) can arrive at a speed no greater than c. Notice that a speed strictly less than c is allowed.

We can make the change of variable $r = c(t-s)$ and change the order of integration. This gives the following result.

Corollary 3.1 *The formula*

$$u(x,t) = \frac{1}{2c^2} \int_0^{ct} \int_{x-r}^{x+r} f(y, t - \frac{r}{c})\, dy\, dr = \frac{1}{2c^2} \int_{x-ct}^{x+ct} \int_{|x-y|}^{ct} f(y, t - \frac{r}{c})\, dr\, dy. \tag{3.20}$$

gives the solution of the inhomogeneous wave equation in one dimension with zero initial condition.

The above formula says that the effect at x of a source at y takes place after a time lag between 0 and $|x-y|/c$. It is interesting to replace the time-dependent source $f(x,t)$ by the static source $c^2 f(x)$. Then the result is

$$u(x,t) = \frac{1}{2} \int_{x-ct}^{x+ct} (ct - |x-y|) f(y)\, dy. \tag{3.21}$$

Does this converge as ct gets large? Only if one subtracts an infinite constant.

3.3 The Kirchoff solution

Now we want to find the solution of the wave equation in three dimensions. The method to be used is that of spherical means. This is just the use of the averages over balls of the form

$$M_{x,r}(h) = \fint_{\partial B(x,r)} h(y)\, dS(y). \tag{3.22}$$

Lemma 3.3 *In n dimensions the spherical means satisfy the partial differential equation*

$$\frac{1}{r^{n-1}} \frac{\partial}{\partial r} r^{n-1} \frac{\partial}{\partial r} M_r(h) = \triangle_x M_{x,r}(h) \tag{3.23}$$

This equation is like the wave equation, except that it has the operator

$$\frac{1}{r^{n-1}}\frac{\partial}{\partial r}r^{n-1}\frac{\partial}{\partial r} = \frac{\partial^2}{\partial r^2} + \frac{n-1}{r}\frac{\partial}{\partial r} \tag{3.24}$$

with an extra first order term. Notice that this operator is the radial part of the Laplace operator.

Proof: Compute

$$r^{n-1}\frac{\partial}{\partial r}M_{x,r}(h) = \frac{1}{n\alpha(n)}\int_{\partial B(x,r)} Dh\cdot\nu\,dS = \frac{1}{n\alpha(n)}\int_{B(x,r)} \triangle h(y)\,dy. \tag{3.25}$$

Then compute

$$\frac{1}{r^{n-1}}\frac{\partial}{\partial r}r^{n-1}\frac{\partial}{\partial r}M_{x,r}(h) = \frac{1}{n\alpha(n)}\int_{\partial B(x,r)} Dh\cdot\nu\,dS = ⨍_{\partial B(x,r)} \triangle h(y)\,dS(y). \tag{3.26}$$

Finally, note that this is

$$⨍_{\partial B(0,r)} \triangle h(x+z)\,dS(z) = \triangle ⨍_{\partial B(0,r)} h(x+z)\,dS(z). \tag{3.27}$$

The reason that this works is that $\triangle h(x+z)$ may be computed as a derivative with respect to either the x or the z variables.

It turns out that this equation works to solve the wave equation in each odd dimension. The simplest and by far the most practically important case is $n = 3$. In this case the key is the operator identity

$$\frac{1}{r^2}\frac{\partial}{\partial r}r^2\frac{\partial}{\partial r} = \frac{1}{r}\frac{\partial^2}{\partial r^2}r. \tag{3.28}$$

This identity tells us that $rM_{x,r}(h)$ satisfies the wave equation in the form

$$\frac{\partial^2}{\partial r^2}rM_{x,r}(h) = \triangle rM_{x,r}(h). \tag{3.29}$$

Theorem 3.2 *In three dimensions the function*

$$u(x,t) = tM_{x,ct}(h) \tag{3.30}$$

is a solution of the homogeneous wave equation with initial condition $u = 0$ *and* $\partial u/\partial t = h$ *at* $t = 0$.

Proof: We have already checked that it is a solution of the wave equation. To check the initial condition, compute

$$\frac{\partial}{\partial t}u(x,t) = M_{x,ct}(h) + t\frac{\partial}{\partial t}M_{x,ct}(h). \tag{3.31}$$

As t approaches zero the first term approaches $h(x)$ and the second term approaches zero. (An even simpler way of getting the same result is to compute the derivative as the limit of $u(x,t)/t = M_{x,ct}(h)$.)

The Kirchoff solution is extremely beautiful. It can of course be written in terms of an integral with respect to surface area as

$$u(x,t) = \int_{\partial B(x,ct)} \frac{h(y)}{4\pi c^2 t}\, dS(y) \tag{3.32}$$

Even more common is the form

$$u(x,t) = \int_{\partial B(x,ct)} \frac{h(y)}{4\pi c|x-y|}\, dS(y). \tag{3.33}$$

There is an elegant expression of the Kirchoff solution as convolution with respect to a certain delta function expression. The solution is the convolution of the initial condition with the function

$$\frac{\delta(r-ct)}{4\pi cr} = \frac{\delta(r^2-(ct)^2)}{2\pi c} \tag{3.34}$$

for $t > 0$. The last expression exhibits the relativistic invariance of the fundamental solution: in each Lorentz frame the solution takes the form of a sphere expanding at the speed c.

The physical meaning of the formula is evident. The value of the solution at x at time t depends only on what was happening at time zero on a sphere about x of radius ct.

One can think of this in another way. If the initial condition $h(y)$ is zero except in a small neighborhood of a point y_0, then the solution $u(x,t)$ is zero except close to the sphere of radius ct about y_0. The solution is an expanding sphere. Before the reader begins to think that this is intuitively obvious, it is wise to mention that such a result is only true in odd dimensions three or more.

We can also solve the inhomogeneous equation by the Duhamel formula

$$u(x,t) = \int_0^t \int_{\partial B(x,c(t-s))} \frac{f(y,s))}{4\pi c|x-y|}\, dS(y)\, ds. \tag{3.35}$$

Then it is natural to make the change of variable $r = c(t-s)$ to change this to an integral over the ball. This gives the following result.

Corollary 3.2 *In three dimensions the formula*

$$u(x,t) = \int_{B(x,ct)} \frac{f(y, t - \frac{|x-y|}{c})}{4\pi c^2|x-y|}\, dy \tag{3.36}$$

gives the solution of the inhomogeneous wave equation with initial condition zero.

The above formula says that the effect at x of a source at y takes place after a time lag of $|x-y|/c$. It is interesting to replace f by $c^2 f$ and then take the limit as c goes to infinity in this last formula. One should recover a well-known result.

3.4 The method of descent

One can try to solve the equation in two dimensions by using the solution already known in three dimensions. One merely has to specialize to an initial condition of a special kind.

Consider the problem of finding the fundamental solution of the wave equation in two dimensions. Let $h(y)$ be the function giving the initial condition, where y ranges over two-dimensional space. Let P be the projection from three dimensional space to two-dimensional space. Then $h(Pz)$ is a function defined for z in three dimensional space. The solution of the wave equation in two dimensions is thus

$$u(x,t) = t\frac{1}{4\pi c^2 t^2}\int_{\partial B(x,ct)} h(Pz)\,dS(z). \tag{3.37}$$

Here x is in two-dimensional space, but the sphere $\partial B(x,ct)$ is the two-sphere in three-dimensional space.

We want to make a change of variable $y = pz$ to reduce this to an integral over the disk $B(x,ct)$ in two-dimensional space. All that we need to do is to find the Jacobian of the change of variable. The result is that

$$u(x,t) = t\frac{1}{4\pi c^2 t^2}2\int_{B(x,ct)} h(y)\frac{1}{\cos(\theta)}\,dy, \tag{3.38}$$

where θ is the angle between the north pole and the point that projects to y. The factor of two comes from the fact that we project from both hemispheres. The cosine factor is

$$\cos(\theta) = \frac{\sqrt{c^2t^2 - (x-y)^2}}{ct}. \tag{3.39}$$

Thus the solution is

$$u(x,t) = t\frac{1}{2\pi ct}\int_{B(x,ct)} h(y)\frac{1}{\sqrt{c^2t^2 - (x-y)^2}}\,dy, \tag{3.40}$$

Again it is t times a probability average. However now the average is over the entire disk $B(x,ct)$, not just the circle $\partial B(x,ct)$, and it is definitely not uniform. The largest contribution comes from near the sphere, where the $1/\cos(\theta)$ factor is largest. But this is really qualitatively quite different from three dimensions, where all the wave propagation is exactly at speed c. In two dimensions some of the solution lingers behind, and we can only say that the maximum speed is c.

3.5 Solution in odd dimension

The identity that is the key to solving the wave equation in odd dimensions greater than three is

$$r^{n-2}[\frac{1}{r^{n-1}}\frac{\partial}{\partial r}r^{n-1}\frac{\partial}{\partial r}]\frac{1}{r^{n-2}} = \frac{\partial^2}{\partial r^2} - \frac{n-3}{r}\frac{\partial}{\partial r}. \tag{3.41}$$

This identity takes a particularly simple form when $n = 3$. When $n > 3$ we need to deal with the extra term on the right hand side.

The extra identity that we need is

$$\frac{\partial^2}{\partial r^2}\left(\frac{1}{r}\frac{\partial}{\partial r}\right) = \left(\frac{1}{r}\frac{\partial}{\partial r}\right)[\frac{\partial^2}{\partial r^2} - \frac{2}{r}\frac{\partial}{\partial r}]. \tag{3.42}$$

Let $n = 3 + 2\ell$. If we apply the previous identity ℓ times, we obtain

$$\frac{\partial^2}{\partial r^2}\left(\frac{1}{r}\frac{\partial}{\partial r}\right)^{\ell} = \left(\frac{1}{r}\frac{\partial}{\partial r}\right)^{\ell}[\frac{\partial^2}{\partial r^2} - \frac{2\ell}{r}\frac{\partial}{\partial r}]. \tag{3.43}$$

Now we combine these two identities. They tell that

$$u(x,t) = C\left(\frac{1}{r}\frac{\partial}{\partial r}\right)^{\ell} r^{2\ell+1}M_{x,r}(h) \tag{3.44}$$

satisfies the homogeneous wave equation

$$\frac{\partial^2}{\partial r^2}u = \triangle u \tag{3.45}$$

in $n = 3 + 2\ell$ dimensions. The proof is to compute

$$\frac{\partial^2}{\partial r^2}u(x,t) = C\left(\frac{1}{r}\frac{\partial}{\partial r}\right)^{\ell} r^{2\ell+1}[\frac{1}{r^{2\ell+2}}\frac{\partial}{\partial r}r^{2\ell+2}\frac{\partial}{\partial r}M_{x,r}(h)]. \tag{3.46}$$

This in turn is

$$C\left(\frac{1}{r}\frac{\partial}{\partial r}\right)^{\ell} r^{2\ell+1}\triangle M_{x,t}(h) = \triangle u(x,t). \tag{3.47}$$

Theorem 3.3 *Let $n = 3 + 2\ell$. Let $\gamma_\ell = 3 \cdot 5 \cdots (2\ell + 1)$. The function*

$$u(x,t) = \frac{1}{\gamma_\ell}\left(\frac{1}{t}\frac{\partial}{\partial t}\right)^{\ell} t^{2\ell+1}M_{x,r}(h) \tag{3.48}$$

is a solution of the homogeneous wave equation

$$\frac{\partial^2}{\partial r^2}u = c^2\triangle u \tag{3.49}$$

with initial condition $u = 0$ and $\partial u/\partial t = h$ at $t = 0$.

Proof: We have already checked that it is a solution of the wave equation. To check the initial condition, compute

$$\left(\frac{1}{t}\frac{\partial}{\partial t}\right)^{\ell} t^{2\ell+1} = \gamma_\ell t. \tag{3.50}$$

Thus the solution u is $tM_{x,ct}(h)$ plus terms with higher powers of t in front. The time derivative $\partial u/\partial t$ is thus $M_{x,ct}(h)$ plus terms with powers of t in front. As t approaches zero this first term approaches $h(x)$ and the other terms each approach zero.

This solution shows that the picture of expanding spheres is true in each odd dimension $n = 3 + 2\ell$ with $\ell = 0, 1, 2, \ldots$. This picture fails in even dimensions. The solutions fill out the ball; part of the solution lags the expanding spherical front. This can be seen again by using the method of descent.

3.6 Conservation of energy

There is an important notion of conservation of energy for the wave equation. Define the energy in the region U to be

$$e(t) = \frac{1}{2}\int_U [\left(\frac{\partial u}{\partial t}\right)^2 + c^2|Du|^2]\,dx. \tag{3.51}$$

Then

$$\frac{de(t)}{dt} = \int_U \frac{\partial u}{\partial t} f\,dx + \int_{\partial U} \frac{\partial u}{\partial t} Du \cdot \nu\,dS. \tag{3.52}$$

This says that the change in energy is due to the force f and to energy flow through the boundary. For the case of a homogeneous equation where $f = 0$ this says that the change in energy in a region is completely accounted for by the flow of energy through the boundary. This of course is the usual form of a conservation law.

This result applies in the case when U is Euclidean space, provided that we assume that the function u and its derivatives approach zero at infinity rapidly enough so that the contribution of the boundary integral may be neglected. For the homogeneous wave equation it says that the energy is constant.

If the energy is constant, then it is equal to the initial value

$$e(0) = \frac{1}{2}\int [h^2 + c^2|Dg|^2]\,dx. \tag{3.53}$$

It is tempting to regard all pairs g and h for which the energy is finite as acceptable initial values for the wave equation. Then the natural result would be that the solution has finite energy equal to this same value for all time. However in order to make this idea precise, we need a more sophisticated notion of derivative. The reason is that the condition of finite energy says merely that h and $|Dg|$ are in L^2, the space of functions with finite square integral. However this does not mean that the functions have to be differentiable at every point. What is needed is a broader concept of derivative, and this is given by the theory of Sobolev spaces. This theory will be presented in a later chapter.

This result is also useful in the case of the wave equation in a bounded region U of space. It gives an important uniqueness theorem.

Theorem 3.4 *Let U be a bounded open set with smooth boundary and let $T > 0$. Let U_T be the space-time cylinder consisting of space time points (x, t) with x in U and $0 < t \leq T$. Let Γ_T be its space-time boundary (which includes the time zero slice but not the time T slice). Consider the homogeneous wave equation*

$$\frac{\partial^2 u}{\partial t^2} = c^2 \triangle u. \tag{3.54}$$

in U_T with initial-boundary condition $u = 0$ on Γ_T and with initial condition $\partial u/\partial t = 0$ on U at $t = 0$. Then $u = 0$ everywhere in U_T.

Proof: Consider the energy integral. Since $u = 0$ on $\partial U \times [0, T]$, it follows that $\partial u/\partial t = 0$ on $\partial U \times [0, T]$. Hence there is no energy flux across the boundary, and so the energy $e(t)$ is constant.

Since $u = 0$ on U at $t = 0$, it follows that $Du = 0$ on U at $t = 0$. Furthermore, $\partial u/\partial t = 0$ on U at $t = 0$. So $e(0) = 0$. It follows that $e(t) = 0$ for each t. It follows that $\partial u/\partial t = 0$ and $Du = 0$ within U_T.

Since $u = 0$ on U at $t = 0$, it follows that $u = 0$ within U_T.

Corollary 3.3 *Let U be a bounded open set with smooth boundary and let $T > 0$. Let U_T be the parabolic cylinder and Γ_T its parabolic boundary. Consider the inhomogeneous wave equation*

$$\frac{\partial^2 u}{\partial t^2} = c^2 \triangle u + f. \tag{3.55}$$

in U_T with initial-boundary condition $u = g$ on Γ_T and with initial condition $\partial u/\partial t = h$ on U at $t = 0$. Then the solution is uniquely specified by g and h.

Chapter 4

Hamilton-Jacobi equation

4.1 Characteristics

The next subject is non-linear first order partial differential equations. We begin however with a linear example. This is the transport equation

$$\frac{\partial u}{\partial t} + \mathbf{b}(x) \cdot Du = c(x)u. \tag{4.1}$$

with initial condition $u(y, 0) = g(y)$.

One important special case of this is the conservation law

$$\frac{\partial u}{\partial t} + \operatorname{div}(\mathbf{b}(x)u) = 0. \tag{4.2}$$

This corresponds to taking $c(x) = -\operatorname{div}(\mathbf{b}(x))$.

We shall see that the solution moves locally with velocity $\mathbf{b}(x)$. If we look at a curve

$$\frac{dx}{dt} = \mathbf{b}(x) \tag{4.3}$$

with initial condition $x = y$ at $t = 0$, the solution is obtained by integrating the differential equation. Such a solution curve is called a characteristic. Here y is the starting point, and x is the corresponding point on the characteristic curve at time t.

Example 1. If $\mathbf{b}$ is constant, then $x = y + t\mathbf{b}$.

Example 2: If $\mathbf{b}(x) = \lambda x$, then $x = ye^{\lambda t}$.

Along this characteristic curve we can use the chain rule to compute that

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + D_x u \cdot \frac{dx}{dt} = c(x)u. \tag{4.4}$$

The final equality follows from the equation for the characteristic curve and the original partial differential equation. The initial condition is $u = g(y)$ at $t = 0$. The solution is obtained by integrating along the curve. The solution is

$$u = g(y)e^{\int_0^t c(x(s))\, ds}. \tag{4.5}$$

To find the solution $u(x,t)$ of the original partial differential equation, solve for y in terms of x and t.

Example 1. For the conservation law with $\mathbf{b}$ constant the solution is

$$u(x,t) = g(x - \mathbf{b}t). \tag{4.6}$$

This describes translational motion at velocity $\mathbf{b}$.

Example 2. For the conservation law with $\mathbf{b}(x) = \lambda x$ the source is $-\lambda \operatorname{div}(x)u = -n\lambda u$, and the solution is

$$u(x,t) = g(xe^{-\lambda t})e^{-n\lambda t}. \tag{4.7}$$

This describes a sort of explosion that spreads everything out.

For the conservation law the integral over space is independent of time. We can see this directly from the partial differential equation. Apply the divergence theorem in the form

$$\frac{d}{dt}\int_V u\,dx + \int_{\partial V} u\mathbf{b}(x)\cdot\nu\,dS = 0. \tag{4.8}$$

If we take the limit as V approaches $\mathbf{R}^n$, and if u approaches zero sufficiently rapidly as x goes to infinity, then the integral of u over space is constant. This says that u is some kind of a density, a conserved quantity.

If the solution is given by the method of characteristics, one can also check the conservation law by direct calculation. The solution obtained by integration is

$$u(x,t) = g(y)\exp(-\int_0^t \operatorname{div}(\mathbf{b}(x(s)))\,ds), \tag{4.9}$$

where the characteristic curve has $x(0) = y$ and $x(t) = x$. It works out that the exponential factor is just the Jacobian determinant $\det(D_x y)$. Therefore the integral of $u(x,t)$ with respect to x is the same as the integral of $g(y)$ with respect to y.

These considerations give the following interpretation of the conservation law. There is a river flowing steadily along with velocity field $\mathbf{b}(x)$. A substance is placed in the river with initial density $u(y,0) = g(y)$. Then $u(x,t)$ is a description of density at later times, when the substance is carried along passively in the river.

We can also look at the transport equation with a source:

$$\frac{\partial u}{\partial t} + \operatorname{div}(\mathbf{b}(x)u) = c(x)u + f(x,t). \tag{4.10}$$

This is solved with the same characteristic curve $dx/dt = \mathbf{b}(x)$. However now $du/dt = c(x)u + f(x,t)$ along the curve.

Example 1. For the conservation law with $\mathbf{b}$ is constant the solution is

$$u(x,t) = g(y - \mathbf{b}t) + \int_0^t f(y - \mathbf{b}(t-s), s)\,ds. \tag{4.11}$$

Example 2. For the conservation law with $\mathbf{b}(x) = \lambda x$ the solution is

$$u(x,t) = e^{-n\lambda t} g(xe^{-\lambda t}) + \int_0^t e^{-n\lambda(t-s)} f(xe^{-\lambda(t-s)}, s)\, ds. \tag{4.12}$$

4.2 Hamiltonian-Jacobi equations

A Hamilton-Jacobi equation (the special kind we are considering) is given by a function H from $\mathbf{R}^n$ to $\mathbf{R}$. The equation is

$$\frac{\partial u}{\partial t} + H(Du) = 0. \tag{4.13}$$

The initial condition is $u(x,0) = g(x)$.

Again this may be solved by the method of characteristics. We need first to guess what the characteristic equations could be. If we expand $H(p)$ around some point p, we obtain $H(p_1) = H(p) + D_pH(p) \cdot (p_1 - p)$ plus higher order terms. Thus we can think of the equation as looking locally like the linear equation

$$\frac{\partial u}{\partial t} + D_pH(p) \cdot Du = p \cdot D_pH(p) - H(p). \tag{4.14}$$

This suggests that the velocity of the characteristic curve should be the quantity $D_pH(p)$. Furthermore, it suggests that the rate of change of the solution along the characteristic curve should be $p \cdot D_pH(p) - H(p)$. Of course we want to rig the situation so that the value of p is actually Du. It is not hard to compute that the derivative of Du along the characteristic curve is zero, so the way to get this is to take p equal to the initial value of Du on the characteristic curve.

For given p define the corresponding velocity of propagation to be

$$q = D_pH(p) \tag{4.15}$$

The first equation for the characteristic curve is

$$\frac{dx}{dt} = q \tag{4.16}$$

with the initial condition $x = y$ when $t = 0$. The other equation for the characteristic curve is

$$\frac{dp}{dt} = 0 \tag{4.17}$$

with the initial condition that $p = Dg(y)$ when $t = 0$. This coupled system of equations has solutions

$$x = y + qt \tag{4.18}$$

and

$$p = Dg(y). \tag{4.19}$$

There may be a problem in determining y as a function of given x and t.

Define

$$L(q) = p \cdot q - H(p) \tag{4.20}$$

The evolution of u along this curve is given by

$$\frac{du}{dt} = L(q) \tag{4.21}$$

with initial condition $u = g(y)$ when $t = 0$. This has solution

$$u = g(y) + tL(q). \tag{4.22}$$

It is not immediately clear that the solution of these equations is also a solution of the original partial differential equation. To prove this, we need a lemma.

Lemma 4.1 *Let $x = y + qt$ with $q = D_pH(p)$ and $p = Dg(y)$, and assume that this defines y locally as a smooth function of x. Let $u(x,t) = g(y) + tL(q)$, where $L(q) = p \cdot q - H(p)$. Then*

$$Du = p. \tag{4.23}$$

Proof: We compute

$$D_xu = p\, D_xy + t[p\, D_xq + q\, D_xp - D_xH(p)]. \tag{4.24}$$

We insert

$$D_xy = I - tD_xq \tag{4.25}$$

and

$$D_xH(p) = q\, D_xp. \tag{4.26}$$

This gives the result.

Theorem 4.1 *Let $x = y + qt$ with $q = D_pH(p)$ and $p = Dg(y)$, and assume that this defines y locally as a smooth function of x. Let $u(x,t) = g(y) + tL(q)$, where $L(q) = p \cdot q - H(p)$. Then*

$$\frac{\partial u}{\partial t} + H(Du) = 0. \tag{4.27}$$

Proof: Compute the derivative of u along the characteristic curve by the chain rule:

$$\frac{du}{dt} = \frac{\partial u}{\partial t} + Du \cdot \frac{dx}{dt} = \frac{\partial u}{\partial t} + p \cdot q, \tag{4.28}$$

by the lemma. On the other hand,

$$\frac{du}{dt} = L(q) = p \cdot q - H(p) = p \cdot q - H(Du), \tag{4.29}$$

again by the lemma. If we compare these two results, we get the theorem.

Example: Our standard example will be the differential equation

$$\frac{\partial u}{\partial t} + \frac{1}{2m}|Du|^2 = 0. \tag{4.30}$$

Here $m > 0$ is a constant. The equation describes the erosion of a surface, where the erosion rate is proportion to the square of the size of the gradient. Steep slopes erode faster.

This example is the choice $H(p) = p^2/(2m)$. Thus the characteristic velocity is $q = p/m$ and the characteristic curve is $x = y + (1/m)Dg(y)t$. Thus the erosion is along the upward gradient of the initial condition. Furthermore, $L(q) = p^2/(2m) = mq^2/2$, and so the solution is $u(x,t) = g(y)+t|Dg(y)|^2/(2m)$. This says that the effect of an initial steep gradient is felt on the higher slopes.

When can we solve for y as a function of x? The implicit function theorem indicates that this can break down when x no longer depends on y, that is, the Jacobian of x with respect to y is singular. This is when the Jacobian of y with respect to x blows up. This Jacobian satisfies

$$I = J + (1/m)D^2g(y)Jt \tag{4.31}$$

Thus we should be in good shape as long as $I + (1/m)D^2g(y)t$ is invertible. This is certainly true when the Hessian $D^2g(y)$ is positive definite. So we can be optimistic when we are in a valley with steeper and steeper sides.

Here is an example. Let $g(y) = (1/2)cy^2$ with $c > 0$. Then the equation for the characteristic curve is $x = y + (ct/m)y$. This can be solved for y; we get $y = x/(1 + ct/m)$. Thus the solution is $u(x,t) = (1/2)c[1 + ct/m]y^2 = (1/2)cx^2/[1 + ct/m]$. This shows the effect of steady erosion. Eventually the valley flattens out.

However if we are in a situation where the higher slopes level off or form a peak, then there can be big trouble. There can be two different values of y with $x = y + (1/m)Dg(y)t$. Which one can we choose?

4.3 The Hopf-Lax solution

Let us return to the general situation. We assume that $H(p)$ is a convex function of p. For simplicity, we assume even that $H(p)$ is uniformly convex. This means that there is a constant $\theta > 0$ such that for each p the Hessian $D_p^2H(p)$ is bounded below by θ. Since for each p the Hessian is a symmetric matrix, this makes sense as a statement about quadratic forms or, equivalently, about eigenvalues. Each eigenvalue is strictly positive, in fact each eigenvalue is bounded below by θ.

We also assume that the equation

$$q = D_pH(p) \tag{4.32}$$

can be solved for p as a function of q. Then we can define the function

$$L(q) = p \cdot q - H(p) \tag{4.33}$$

Then it is easy to see that

$$p = D_q L(q), \tag{4.34}$$

so the relation between $H(p)$ and $L(q)$ is symmetric. We can compute the Hessian $D_q^2 L(q)$ and see that it is the inverse of the Hessian $D_p^2 H(p)$. Therefore the eigenvalues of $D_q^2 L(q)$ are the reciprocals of the eigenvalues of $D_p^2 H(p)$. Thus the eigenvalues of $D_q^2 L(q)$ are also strictly positive. Furthermore, for all q they are bounded above by the constant $1/\theta$.

Example: In the example above, the Hessian $D_p^2 H(p)$ is $1/m$ times the identity matrix, while the Hessian $D_q^2 L(q)$ is m times the identity matrix.

The Hopf-Lax formula is a proposal for a solution of the Hamilton-Jacobi equation that works even when the method of characteristics gives an ambiguous solution or no solution at all.

We want to define a more general solution for which

$$u(x,t) = g(y) + tL(q), \tag{4.35}$$

where

$$x = y + tq. \tag{4.36}$$

The idea is to eliminate q from this system of equations. This makes the solution depend only on y. The Hopf-Lax formula says to take the y in $\mathbf{R}^n$ for which the solution is minimal:

$$u(x,t) = \min_y [g(y) + tL(\frac{x-y}{t})]. \tag{4.37}$$

This formula makes sense even if there are points at which g is not differentiable.

The power of the Hopf-Lax formula is that it makes sense in a very general situation when the Hamilton-Jacobi equation is ambiguous. However it is related to the Hamilton-Jacobi equation in the following sense.

Theorem 4.2 *Let $u(x,t)$ be given by the Hopf-Lax formula. Assume that the minimum is assumed at a point where the derivative is zero. Assume furthermore that this minimum is unique. Then the solution satisfies the Hamilton-Jacobi equation.*

Proof: The condition that the derivative is zero is that

$$D_y g(y) = D_q L(\frac{x-y}{t}). \tag{4.38}$$

Let y be the point where the derivative is zero. By the assumption that the minimum is unique, this defines y locally as a smooth function of x and t. Define q by requiring that $x = y + qt$. Define $p = D_q L(q)$. Then the condition that the derivative is zero is that $p = Dg(y)$. Furthermore, $u = g(y) + tL(q)$. So this is one of the solutions given by the method of characteristics, with a particular choice of y.

Remark 1. In general there will be n dimensional surfaces in the $n+1$ dimensional space-time where the Hamilton-Jacobi equation is not satisfied by the Hopf-Lax formula. The reason is that there will be distinct points y_a and y_b where the minimum is assumed. The value $u(x,t) = g(y_a) = g(y_b)$ will be the same, but there can be a jump between the values $Dg(y_a) \neq Dg(y_b)$. So there is no reason to expect the gradient $D_x u(x,t)$ to be defined at such points x and t.

Remark 2. This shows that it is natural to take initial conditions for the Hamiltonian-Jacobi equation that have Dg not defined on some $n-1$ dimensional surface in $\mathbf{R}^n$. In this case the Hopf-Lax solution may arise from a minimum point y where the derivative does not exist.

These remarks indicate that the Hopf-Lax formula provides only a weak solution of the Hamiltonian-Jacobi equation, that is, a solution that is valid at most points in space-time but not at all points.

Example: In the model of erosion, the Hopf-Lax solution is

$$u(x,t) = \min_y [g(y) + \frac{m}{2}\frac{(x-y)^2}{t})]. \tag{4.39}$$

For each y the function in brackets is a parabola. The solution is thus a minimum of a family of parabolas parameterized by y.

Let us take the initial condition $g(y) = (1/2)cy^2$ for $|y| \leq a$ and $g(y) = (1/2)ca^2$ for $|y| \geq a$. This is a valley formed in the middle of a plain. The equation for the characteristics is $x = y + (ct/m)y$ for $|y| \leq a$ and $x = y$ for $|y| \geq a$. Clearly there is an ambiguity about what y goes with a given x.

This ambiguity is resolved by the Hopf-Lax solution. The function to be minimized is $g(y) + m(x-y)^2/(2t)$. The minimum occurs either where $x = [1+ct/m]y$ or where $x = y$. It occurs at the first point when $x^2/[1+ct/m] < a^2$. Then the solution is $u(x,t) = (1/2)cx^2/[1+ct/m]$. If $|x|$ is larger than this value, it occurs at the second point, and the solution is $u(x,t) = (1/2)ca^2$. This shows that the valley eats into the plain at a linear rate.

4.4 Weak solutions

We have seen that even if the initial conditions are smooth, it may be that the characteristic curves will cross. This can happen when the initial condition is not convex. In this case the Hopf-Lax solution will have slope discontinuities. Thus there will be surfaces in space-time for which the original partial differential equation is not satisfied. However once this is admitted, then it is no longer clear that the partial differential equation and the initial condition determine the solution.

Example: Take the model of erosion

$$\frac{\partial u}{\partial t} + \frac{1}{2m}|Du|^2 = 0. \tag{4.40}$$

Let $u(x,t) = |x| - t/(2m)$ for $|x| \leq t/(2m)$ and $u(x,t) = 0$ otherwise. This describes the spontaneous formation of a valley. The partial differential equation is satisfied everywhere except on the cone (expanding sphere) $|x| = t/(2m)$. Furthermore, it is zero at $t = 0$. However this solution is clearly not meaningful.

The paradox is that the mere fact of being a solution of the Hamiltonian-Jacobi equation at most points is not enough. The solutions given by the Hopf-Lax method must have some extra property. The property is roughly that the second derivative D^2u must be bounded above by a constant. That is, a sharp peak pointing down is forbidden. However there is nothing ruling out a sharp peak pointing upward, and these are indeed produced by the Hopf-Lax solution. We shall see that the precise condition is that the solution $u(x,t)$ be semiconcave as a function of x.

The condition that a function f is semiconcave is that there is a constant C such that for each point x and unit vector z the second differences

$$\frac{f(x+az) - 2f(x) + f(x-az)}{a^2} \leq C. \tag{4.41}$$

Note that the limit as a tends to zero of the left hand side would be the derivative $z \cdot D^2 f(x) z$, if the limit existed. (The terminology semiconcave is perhaps confusing; the condition is really that the function not be excessively convex.)

When the limit exists, the situation is simple. The condition that $z \cdot D^2 f(x) z \leq C$ for unit vectors z is enough to imply that f is semiconcave. This follows from a Taylor expansion of $f(x \pm az)$ with second order remainder.

Lemma 4.2 *Let f_y be a family of semiconcave functions with constant C for y running through a parameter set. Assume that for each x the minimum $m(x) = \min_y f_y(x)$ exists. Then the function $m = \min_y f_y$ is also semiconcave with constant C.*

Proof: Fix x. For this x, there exists y such that $m(x) = f_y(x)$. Thus

$$m(x+az) - 2m(x) + m(x-az) = m(x+az) - 2f_y(x) + m(a-az). \tag{4.42}$$

Furthermore, for this y and all x' we have $m(x') \leq f_y(x')$. So this in turn is bounded by

$$f_y(x+az) - 2f_y(x) + f_y(x+az) \leq Ca^2. \tag{4.43}$$

Theorem 4.3 *The Hopf-Lax formula defines a function $u(x,t)$ that is semiconcave as a function of x for all $t > 0$.*

Proof: The uniform convexity assumption on $H(p)$ implies that the $D^2L(q)$ is bounded above by a constant independent of q. Thus $L(q)$ is semiconcave with some constant C. It follows that for each y the function $g(y) + tL(\frac{x-y}{t})$ is semi-concave with constant C/t. Since the Hopf-Lax solution is

$$u(x,t) = \min_y [g(y) + tL(\frac{x-y}{t})], \tag{4.44}$$

it follows that the same is true for $u(x,t)$.

Example: Take the erosion problem with $H(p) = p^2/(2m)$ with initial condition $g(y) = c|y|$. Here $m > 0$ and $c > 0$ and $L(q) = (m/2)q^2$. This is an artificially created valley with a sharp bottom. The initial condition is not smooth, and in fact it is not even semiconcave. The characteristic velocity for the characteristic starting at y is $(c/m)y/|y|$. The characteristic starting at y is $x = y + (ct/m)\, y/|y|$. The solution for y in terms of x is $y = x - (ct/m)\, x/|x|$. This does not work for the region of space time with $|x| < ct/m$, which is not even reached by characteristics. However the Hopf-Lax solution still makes sense. The minimum of $c|y| + (m/2)(y - x)^2/t$ occurs for y belonging to a characteristic or for $y = 0$. The corresponding values are $c|x| - c^2t/(2m)$ and $(m/2)x^2/t$. The former is the minimum for $|x| \geq ct$ and the latter is the minimum for $|x| \leq ct$. The sharp valley bottom is smoothed out by the erosion process. The resulting profile is semiconcave for $t > 0$.

The general discussion of this section shows that the Hopf-Lax formula defines a weak solution of the Hamilton-Jacobi equation that is semiconcave. Evans gives a converse to this result. It says that a weak solution of the Hamilton-Jacobi equation that is semiconcave must be the Hopf-Lax solution.

4.5 The Hopf-Cole transformation for the quadratic Hamilton-Jacobi equation

If we take a solution $v > 0$ of the heat equation

$$\frac{\partial v}{\partial t} = \frac{1}{2}\sigma^2 \triangle v \tag{4.45}$$

and make the change of variables

$$v = \exp(-\frac{1}{m\sigma^2}u), \tag{4.46}$$

then we obtain the viscous quadratic Hamilton-Jacobi equation

$$\frac{\partial u}{\partial t} + \frac{1}{2m}|Du|^2 = \frac{1}{2}\sigma^2 \triangle u. \tag{4.47}$$

This is called the Hopf-Cole transformation.

The viscous equation describes the erosion process, but now there is an extra term that says that valleys erode slower, while peaks erode particularly rapidly.

We can solve this equation by reducing it to the heat equation. The result is that

$$u(x,t) = -m\sigma^2 \log(v(x,t)), \tag{4.48}$$

where

$$v(x,t) = \int (2\pi\sigma^2 t)^{-\frac{n}{2}} \exp(-\frac{(x-y)^2}{2\sigma^2 t}) \exp(-\frac{g(y)}{m\sigma^2})\, dy. \tag{4.49}$$

The interesting thing is that we may take the limit as σ^2 tends to zero. This is a simple application of Laplace's method. Fix x and t. Assume that there is only one point $y_{x,t}$ for which

$$K(x, y, t) = \frac{(x-y)^2}{2t} + \frac{g(y)}{m} \tag{4.50}$$

is minimal. Then by Laplace's method

$$-\sigma^2 \log(v(x,t)) \to K(x, y_{x,t}, t) \tag{4.51}$$

as $\sigma^2 \to 0$. Notice that since $\sigma^2 \log(\sigma^2) \to 0$ in this limit, the prefactors involving powers of σ^2 give no contribution. This is the same thing as saying that the solution in the limit $\sigma^2 \to 0$ is

$$u(x,t) = mK(x, y_{x,t}, t) = \frac{m(x - y_{x,t})^2}{2t} + g(y_{x,t}) \tag{4.52}$$

where y is minimal. This result is indeed the Hopf-Lax solution.

4.6 Laplace's method

Here we give the lowest order estimates for the asymptotics of an integral with a parameter β with an integrand $\exp(-\beta h(y))$. We review the theory without giving full proofs.

Theorem 4.4 *Assume that h is a continuous function that is bounded below and that grows at least as fast as a linear function near infinity. Suppose that there is a unique point y_0 at which $h(y)$ has a minimum. Let*

$$Z(\beta) = \int \exp(-\beta h(y))\, dy. \tag{4.53}$$

and consider the probability density

$$\frac{1}{Z(\beta)} \exp(-\beta h(y)). \tag{4.54}$$

Let f be a continuous function whose magnitude grows no faster than a polynomial. Let

$$E_\beta(f) = \frac{1}{Z(\beta)} \int f(y) \exp(-\beta h(y))\, dy \tag{4.55}$$

be the expectation of f with respect to this probability density. Then

$$E_\beta(f) \to f(y_0) \tag{4.56}$$

as $\beta \to \infty$.

Proof: As $\beta \to \infty$ the probability density becomes more and more concentrated near the point y_0.

Theorem 4.5 *Assume that h is a continuous function that is bounded below and that grows at least as fast as a linear function near infinity. Suppose that there is a unique point y_0 at which $h(y)$ has a minimum. Let*

$$Z(\beta) = \int \exp(-\beta h(y))\, dy. \tag{4.57}$$

Then

$$-\frac{1}{\beta} \log Z(\beta) \to h(y_0) \tag{4.58}$$

as $\beta \to \infty$.

Proof: Since $-\log Z(\beta)$ and β are both going to infinity, by l'Hospital's rule the limit is the same as the limit of

$$-\frac{d}{d\beta} \log(Z(\beta)) = E_\beta(h). \tag{4.59}$$

Apply the previous theorem.

Sometimes the result of this last theorem is written in the form

$$Z(\beta) \sim \exp(-\beta h(y_0)) \tag{4.60}$$

as $\beta \to 0$. What this means is what the theorem states:

$$-\lim_{\beta\to\infty} \frac{1}{\beta} \log(Z(\beta)) = h(y_0). \tag{4.61}$$

It is perhaps surprising that the result does not depend on any other details of the integration. However taking the logarithm washes out a lot of details.

Remark: The notation in this section is that of statistical mechanics. In that subject h is the energy function and β is the reciprocal of the temperature. The probability density is called the Gibbs density. The first result says that the expectation of f at zero temperature comes from the state with minimum energy. The quantity $F(\beta) = -\log(Z(\beta))/\beta$ is called the free energy. This definition is more intuitive if one writes it in the form

$$\exp(-\beta F(\beta)) = \int \exp(-\beta h(y))\, dy, \tag{4.62}$$

because it exhibits the relation of the free energy to an integral involving the energy. The second result says that the free energy at zero temperature is the minimum energy.

Chapter 5

Conservation laws

5.1 Scalar conservation laws

A conservation law (again a special kind) is given by a function $\mathbf{F}$ from $\mathbf{R}$ to $\mathbf{R}^n$. The equation is

$$\frac{\partial u}{\partial t} + \operatorname{div} \mathbf{F}(u) = 0. \tag{5.1}$$

The initial condition is $u(x, 0) = g(x)$. This equation may also be written in the form

$$\frac{\partial u}{\partial t} + \mathbf{F}'(u) \cdot Du = 0. \tag{5.2}$$

Thus it is quasi-linear; the derivatives of u occur only to the first power.

This may also be solved by the method of characteristics. The characteristic curves depend on the solution. The curve is the solution of

$$\frac{dx}{dt} = \mathbf{F}'(u), \tag{5.3}$$

with initial condition $x = y$ when $t = 0$. The solution is $x = y + \mathbf{F}'(u)t$. It is to be obtained by solving the equation $x = y + \mathbf{F}'(g(y))t$ for y as a function of x and t. This is an implicit equation, so there might be a problem with finding a unique solution.

Along the solution curve the solution satisfies

$$\frac{du}{dt} = 0. \tag{5.4}$$

with initial condition $u = g(y)$ at $t = 0$. Thus it is constant with this value along the curve. The solution is thus obtained by finding the y for the start of the characteristic curve that arrives at x, t and using $u = g(y)$.

The solution could also be obtained equivalently by solving the equation $u = g(x - \mathbf{F}'(u)t)$ for u as a function of x and t. This is also an implicit equation, so again there can be difficulties in finding a global solution.

5.2 Conservation in one space dimension

In one dimension the conservation law involves a function F from $\mathbf{R}$ to $\mathbf{R}$. The equation in conservation form is

$$\frac{\partial u}{\partial t} + \frac{\partial F(u)}{\partial x} = 0. \tag{5.5}$$

The initial condition is $u(x, 0) = g(x)$. It would seem evident that the equation may also be written in the form

$$\frac{\partial u}{\partial t} + F'(u) \cdot \frac{\partial u}{\partial x} = 0, \tag{5.6}$$

and this is indeed true for classical solutions. However we shall see that we need to consider more general weak solutions for which this form of the equation is ambiguous. The conservation form is more fundamental.

For classical solutions we use the method of characteristics. The characteristic curves are the solutions of

$$\frac{dx}{dt} = F'(u). \tag{5.7}$$

The solution is $x = y + F'(u)t$. Along the solution curve the solution satisfies

$$\frac{du}{dt} = 0. \tag{5.8}$$

with initial condition $u = g(y)$ at $t = 0$. Thus it is constant with this value along the curve. It follows that the solution is obtained by solving the equation $x = y + F'(g(y))t$ for y as a function of x and t. The solution is then $u = g(y)$.

Example. Take $F(u) = (1/2)u^2$. This is the Burgers equation. The physical interpretation is that u is the velocity of a gas of free particles. The equation says that the acceleration of the gas is zero. The equation for characteristics is $x = y + ut$, that is, $x = y + g(y)t$. Once we have solved this equation for y as a function of x and t, then we have the solution $u(x, t) = g(y)$. If we compute $dx/dy = 1 + g'(y)t$, then we see that if $g'(y) > 0$, then x and y increase together. So there is hope for a well-defined solution. On the other hand, if $g'(y) < 0$, then the fast particles eventually catch up with the slow particles, and there is an ambiguity.

5.3 The Lax-Oleinik solution

We now assume that $F(p)$ is a convex function of p and that $F(0) = 0$. Consider the Hamiltonian-Jacobi equation

$$\frac{\partial w}{\partial t} + F(\frac{\partial w}{\partial x}) = 0 \tag{5.9}$$

with the initial condition $w = h$ when $t = 0$. Then the solution of the conservation law is

$$u = \frac{\partial w}{\partial x}. \tag{5.10}$$

The initial condition is that $u = g$ when $t = 0$, where $g = \partial h/\partial x$.

We have the Hopf-Lax solution of the Hamiltonian-Jacobi equation:

$$w(x,t) = \min_y [h(y) + tL(\frac{x-y}{t})]. \tag{5.11}$$

Here $L(q) = pq - F(p)$, where $q = F'(p)$. We want the solution u of the conservation law to satisfy $x = y + F'(u)t$. This gives the idea for the solution of the conservation law.

The Lax-Oleinik solution of the conservation law

$$\frac{\partial u}{\partial t} + \frac{\partial F(u)}{\partial x} = 0 \tag{5.12}$$

with initial condition is $u(x,0) = g(x)$ is defined as follows. Let $h(x) = \int_0^x g(y)\,dy$. Let y be the point at which the minimum of

$$h(y) + tL(\frac{x-y}{t}) \tag{5.13}$$

is assumed, where $L(q) = pq - F(p)$ and p is defined by solving $q = F'(p)$. Then the solution u at x and t is defined by solving

$$x = y + F'(u)t. \tag{5.14}$$

The method may be summarized as follows. Fix x and t. For each possible initial point y, define u by $x = y + F'(u)t$. Then minimize $h(y) + tL(u) = h(y) + t[uF'(u) - F(u)]$ with respect to y. If g is continuous at this point, then the fact that the derivative is zero gives $g(y) = u$. So this is a reasonable notion of solution.

Example: Take the example of the Burgers equation. Then the problem is to minimize $h(y) + tu^2/2$, where u is defined as a function of y by $x = y + ut$. As x varies continuously, it is possible that the y where the minimum is assumed makes a jump. Then the solution will be discontinuous as a function of x. This can happen when $g(y)$ is decreasing, so $h(y)$ is concave. For large t the function $h(y) + (y-x)^2/(2t)$ may have a local maximum with a local minimum on either side. As x varies, the absolute minimum jumps from one local minimum to the other.

5.4 Distributions

The basic notion of distribution or weak solution is simple. Suppose that one has an expression involving derivatives of a function $f(x)$. For example, suppose that $df(x)/dx = g(x)$. Then we can multiply by a smooth function v with compact support and integrate and get

$$\int_{-\infty}^{\infty} \frac{d}{dx} f(x)\, v(x)\, dx = \int_{-\infty}^{\infty} g(x) v(x)\, dx. \tag{5.15}$$

Now integrate by parts. This gives

$$-\int_{-\infty}^{\infty} f(x)\, \frac{d}{dx} v(x)\, dx = \int_{-\infty}^{\infty} g(x) v(x)\, dx. \tag{5.16}$$

If this last equation is satisfied for all such test functions v, then we say that $df(x)/dx = g(x)$ in the sense of distributions. This can be true even if $f(x)$ is not differentiable in the classical sense.

Sometimes we may want to write an equation such as $df(x)/dx = \delta(x)$, where $\delta(x)$ is the Dirac delta function. This delta function is actually a shorthand for measure that evaluates the test functions at the point zero. Thus the rigorous meaning of this formula is that

$$-\int_{-\infty}^{\infty} f(x)\, \frac{d}{dx} v(x)\, dx = v(0) \tag{5.17}$$

for all test functions v. An example of a solution of this equation is $f(x) = H(x)$, where H is the Heaviside function defined by $H(x) = 1$ for $x > 0$ and $H(x) = 0$ for $x < 0$. Then the above equation says that

$$-\int_{0}^{\infty} \frac{d}{dx} v(x)\, dx = v(0), \tag{5.18}$$

which is obviously true.

Here is an example of a typical distribution calculation. We claim that for each smooth function we have the product formula $d/dx(g(x)H(x)) = g'(x)H(x) + g(x)\delta(x)$. The proof is to integrate both sides with a test function $v(x)$. This shows that the formula is equivalent to the formula $-\int_{-\infty}^{\infty} g(x)H(x)v'(x)\, dx = \int_{-\infty}^{\infty} g'(x)H(x)v(x)\, dx + g(0)v(0)$. This is the same as $-\int_0^{\infty} g(x)v'(x)\, dx = \int_0^{\infty} g'(x)v(x)\, dx + g(0)v(0)$. Since v has compact support, this follows from ordinary integration by parts.

With some care these ideas can be applied to conservation laws. Suppose that one has a classical solution of the equation

$$\frac{\partial u}{\partial t} + \frac{\partial F(u)}{\partial x} = 0 \tag{5.19}$$

with $u(x, 0) = g(x)$. Let v be a smooth function of x and t that vanishes for large x and for large positive t. Multiply the equation by v and integrate over space and over positive time. Then integrate by parts. This gives

$$-\int_0^{\infty} \int_{-\infty}^{\infty} u \frac{\partial v}{\partial t}\, dx\, dt - \int_0^{\infty} \int_{-\infty}^{\infty} F(u) \frac{\partial v}{\partial x} - \int_{-\infty}^{\infty} g(x) v(x, 0)\, dx = 0. \tag{5.20}$$

The last term comes is the boundary term at $t = 0$ from the integration by parts.

A function u that satisfies the above equation for all smooth v is called a distribution solution or weak solution of the equation. It is possible that such a distribution solution has discontinuities, so the process cannot be reversed to get the original classical equation. So this gives a more general notion of solution. Sometimes we say that the original equation is satisfied, but only in the weak sense or in the sense of distributions.

It is usually impossible to perform nonlinear operations in the context of distribution theory. Therefore for the notion of distribution solution of such a nonlinear equation it is essential that the equation be written in conservation form, so that the nonlinear expression $F(u)$ is evaluated before the derivative is taken. That is, the linear equation

$$\frac{\partial u}{\partial t} + \frac{\partial J}{\partial x} = 0 \tag{5.21}$$

is taken as a distribution equation. The individual terms in this equation can involve delta function. On the other hand, the non-linear relation

$$J = F(u) \tag{5.22}$$

must be taken in the classical sense in which u and J are functions.

5.5 Weak solutions

The solution of the Hamiltonian Jacobi equation may have slope discontinuities, but it satisfies

$$\frac{\partial w}{\partial t} + F(\frac{\partial w}{\partial x}) = 0 \tag{5.23}$$

at most points, and the derivatives may be discontinuous, but they are at least functions. However the solution $u = \partial w/\partial x$ of the conservation law may have actual discontinuities. Therefore its partial derivatives must be taken in the sense of distributions.

If we take the derivative of the equation above in the sense of distributions, we get

$$\frac{\partial u}{\partial t} + \frac{\partial F(u)}{\partial x} = 0. \tag{5.24}$$

The individual terms in this equation may involve delta functions, but they must cancel out.

Let us see what this means in a more computational way. Consider a curve $x = s(t)$ on which the solution has a discontinuity. Such a discontinuity is called a shock. The solution may be written

$$u(x,t) = u_\ell(x,t)H(s(t)-x) + u_r(x,t)H(x-s(t)). \tag{5.25}$$

Thus we have one solution on the left where $x < s(t)$ and another on the right where $x > s(t)$. Correspondingly, we have

$$F(u(x,t)) = F(u_\ell(x,t))H(s(t)-x) + F(u_r(x,t))H(x-s(t)). \tag{5.26}$$

The partial differential equation then gives a condition involving delta functions. Plug the above expressions into the equation and differentiate using the product rule. The result is a delta function $\delta(x - s(t))$ multiplied by a coefficient. For the equation to be satisfied, this coefficient must be zero. Thus we obtain

$$u_\ell(s(t),t)s'(t) - u_r(s(t),t)s'(t) - F(u_\ell(s(t),t)) + F(u_r(s(t),t)) = 0. \tag{5.27}$$

This equation may be written in brief as

$$[F(u_\ell) - F(u_r)] = s'[u_\ell - u_r]. \tag{5.28}$$

This relates the velocity of the shock to the magnitude of the jump.

Example. We know that with an initial condition that is decreasing in space a shock will eventually form. So for simplicity, let us look at the case when a shock is present at the outset. Take the Burgers equation with $F(u) = u^2/2$. We think of u as being the velocity of a gas of particles. The initial condition is that $u = 1$ for $y < 0$ and $u = 0$ for $y > 0$. The fast particles are behind the slow particles. The characteristics on the left are $x = y + t$, and the characteristics on the right are $x = y$. The corresponding solutions are $u = 1$ and $u = 0$. Obviously, these are not compatible. The shock velocity is $s' = 1/2$. Therefore the shock is the line $x = t/2$. To the left of this line the solution is 1, to the right the solution is 0. This says that the fast particles catch up with the slow particles along the shock. This slows down the fast particles and speeds up the slow particles. This is exactly the same solution that is obtained from the Lax-Oleinik principle, where one minimizes $w = (x-y)^2/(2t) + h(y)$, with $h(y) = y$ for $y < 0$ and $h(y) = 0$ for $y > 0$. The minimum values in the two regions are where $y = x - t$ and $y = x$, and the corresponding values of w are $w = t/2+y = x-t/2$ and $w = 0$. The crossover is indeed along the line $x = t/2$.

Example. Take again the Burger's equation. This time take the slow particles behind the fast particles. The initial condition is that $u = 0$ for $y < 0$ and $u = 1$ for $y > 0$. The characteristics on the left are $x = y$, and the characteristics on the right are $x = y + t$. The corresponding solutions are $u = 0$ and $u = 1$. The solution is not defined by the method of characteristics in the region $0 < x < t$. However we can try for a shock solution. Again the shock velocity is $1/2$. So one could expect a shock along the line $x = t/2$ again, with the slow particles being tugged along by the fast particles. The characteristics to the left have velocity zero, the characteristics to the right have velocity one. It is easy to check that this is a weak solution of the partial differential equation.

However there is another solution to this same problem, in which there is no shock. For this solution, it is as if the missing characteristics emerge from the origin. The solution in the region $0 < x < t$ is given by $u = x/t$. This says that the particles in the intermediate region have intermediate velocities. It is easy to check that this is a solution of the partial differential equation. In fact, it is the Lax-Oleinik solution. We need to minimize $w = (x-y)^2/(2t) + h(y)$, where $h(y) = 0$ for $y < 0$ and $h(y) = y$ for $y > 0$. The values in the two regions are $y = x$ for $x < 0$ and $y = x - t$ for $x > t$. There is a third possible value corresponding to the point $y = 0$ where the derivative does not exist. The three corresponding values of w are $w = 0$, $w = t/2 + y = x - t/2$, and $w = x^2/(2t)$. When $0 < x < t$ the only meaningful solution is the third one. The solution in this region is given by solving $x = ut$ for u. In the other regions the other solutions are meaningful and give smaller values of w.

5.6 Entropy

The condition that a function solves the equation in the sense of distributions is not sufficiently restrictive. One wants to find some extra criterion that single out the solutions that are meaningful. This is given by the entropy condition. This says that the function cannot increase faster than linearly as one moves to the right. In particularly, it cannot jump up. However it most certainly can jump down.

Physically this says that the fast moving particles can catch up with the slow moving particles and stick together to form a shock. However they cannot separate and fly apart.

The entropy condition may be derived from the semiconcavity condition discussed in connection with Hamilton-Jacobi equation. Recall that a function is semiconcave if there is a constant with the second difference

$$\frac{f(x+a) - 2f(a) - f(x-a)}{a^2} \leq C. \tag{5.29}$$

We can also write this as

$$[f(x+a) - f(x)] - [f(x) - f(x-a)] \leq Ca^2. \tag{5.30}$$

Apply this formula to the points $x + ka$ for $k = 1$ through n, and then add. The series telescopes, and one gets

$$[f(x+na+a) - f(x+na)] - [f(x+a) - f(x)] \leq Cna^2. \tag{5.31}$$

This can also be written as

$$\frac{1}{y}\left[\frac{f(x+y+a) - f(y)}{a} - \frac{f(x+a) - f(x)}{a}\right] \leq C, \tag{5.32}$$

where $y = na$. If f is differentiable, this shows that

$$\frac{f'(x+y) - f'(x)}{y} \leq C. \tag{5.33}$$

The derivative can only grow linearly.

Note that this fact is obvious if one formulate the semi-concavity condition as $f''(x) \leq C$. However it is nice that one does not need to assume that the second derivative exists.

In the case of the Hamilton-Jacobi equation we have that the solution w satisfies the semiconcavity condition with constant C/t. Since the solution u of the conservation law is given at most points by $u = \partial w/\partial x$, it seems plausible that the Lax-Oleinik solution of the conservation law u satisfies the entropy condition

$$\frac{u(x+y,t) - u(x,t)}{y} \leq \frac{C}{t}. \tag{5.34}$$

It might be better to derive this result directly from the Lax-Oleinik solution, and this can be done. However it is also good to see the direct relation to semiconcavity.

The implications of the entropy condition for a shock is that the solution cannot jump up. In other words, the solution on the right of the discontinuity cannot exceed the solution on the left:

$$u_\ell \geq u_r. \tag{5.35}$$

The entropy condition shows that while the solutions of the conservation law may be discontinuous, they cannot be too irregular for $t > 0$. To see this, let $k > C/t$ and consider the function $v(x,t) = u(x,t) - kx$ as a function of x. It is easy to check that $v(x+y,t) - v(y,t) = u(x+y,t) - u(x,t) - ky \leq (C/t)y - ky < 0$ for $y > 0$. Therefore $v(x,t)$ is a decreasing function of x. It follows that $u(x,t) = v(x,t) + kx$ is the sum of a decreasing function and a linear function. Therefore, the only discontinuities of $u(x,t)$ as a function of x are countably many jump discontinuities.

Example: Here is an example to show how terrible weak solutions can be when they are not restricted by the entropy condition. Consider the Burgers equation in which $F(u) = u^2/2$. Take the initial condition to be identically zero. Any reasonable solution should remain identically zero. But there is nevertheless a non-zero solution of the conservation law. Such a solution is to take $u = -1$ for $-t/2 < x < 0$ and $u = 1$ for $0 < x < t/2$ and $u = 0$ elsewhere. This represents a spontaneous explosion of particles. There are quite reasonable shocks along the lines $x = \pm t/2$. However there is also a shock along the line $x = 0$ that violates the entropy condition.

5.7 Entropy increase

Next we can look at entropy for the conservation law. We will take ϕ to be a convex function. The interpretation is going to be that $-\phi(u)$ is the entropy. We take Y to be another functions satisfying $Y'(u) = \phi'(u)F'(u)$. The interpretation is going to be that $-Y(u)$ is the entropy flux. This should give a conservation law for entropy in the case of smooth solutions. Indeed, it is easy to check that

$$\frac{\partial \phi(u)}{\partial t} + \frac{\partial Y(u)}{\partial x} = \phi'(u)\frac{\partial u}{\partial t} + \phi'(u)F'(u)\frac{\partial u}{\partial x} = 0 \tag{5.36}$$

for a smooth solution.

However things are more interesting when we look at distribution solutions. Then the above calculation is not even meaningful, since we are multiplying the distribution derivative $\partial u/\partial x$ by the function $F'(u)$, which may be not at all smooth.

We want to argue that for such solutions, at least in certain cases, we have

$$\frac{\partial \phi(u)}{\partial t} + \frac{\partial Y(u)}{\partial x} \leq 0. \tag{5.37}$$

This says that in the presence of shocks entropy $-\phi(u)$ need not be conserved. It fact, it increases.

Let us calculate the left hand side at a shock. The coefficient of the delta function is

$$[\phi(u_\ell) - \phi(u_r)]s' - [Y(u_\ell) - Y(u_r)]. \tag{5.38}$$

Here s' is the velocity of the shock, which is determined by the condition that the original conservation law is satisfied:

$$[u_\ell - u_r]s' - [F(u_\ell) - F(u_r)] = 0. \tag{5.39}$$

The following theorem refers to the entropy defined by $-\phi(u)$ with a convex function ϕ. It says that the entropy increases at weak shocks. This is a consequence of the previous entropy condition that says that the solution is larger on the left than on the right of the shock.

Theorem 5.1 *Consider a conservation law with nonlinearity $F(u)$. Consider a solution with a shock that jumps from u_ℓ to u_r. Assume that $F''(u_r) > 0$ and $\phi''(u_r) > 0$. If $u_\ell - u_r > 0$ is small enough, then the delta function source term in the equation for negative entropy $\phi(u)$ at the shock has coefficient*

$$[\phi(u_\ell) - \phi(u_r)]s' - [Y(u_\ell) - Y(u_r)] < 0, \tag{5.40}$$

where s' is the shock velocity. Thus the entropy increases across a weak shock.

Proof: We will think of these functions as depending on $u = u_\ell - u_r$ with fixed u_r. We may also incorporate the constant terms into the definitions of the functions. Thus we now have functions $F(u)$, $\phi(u)$, and $Y(u)$ that each vanish at $u = 0$. Furthermore $F''(0) > 0$ and $\phi''(0) > 0$. The functions are related by $Y'(u) = \phi'(u)F'(u)$. Furthermore, $F(u) = us(u)$. We need to prove that $\phi(u)s(u) - Y(u) < 0$ for small $u > 0$.

As a preliminary step we differentiate $F(u) = us(u)$ and get $F'(u) = us'(u) + s(u)$, $F''(u) = us''(u) + 2s'(u)$, $F'''(u) = us'''(u) + 3s''(u)$. In particular, $F'(0) = s(0)$, $F''(0) = 2s'(0)$, and $F'''(0) = 3s''(0)$.

The main task is to obtain information about $\phi(u)s(u) - Y(u)$. This clearly vanishes at zero.

First we differentiate $\phi(u)s(u) - Y(u)$. The result is $\phi'(u)s(u) + \phi(u)s'(u) - Y'(u) = \phi'(u)s(u) + \phi(u)s'(u) - \phi'(u)F'(u)$. The value at zero of this first derivative is $\phi'(0)s(0) - \phi'(0)F'(0) = 0$.

Differentiate a second time. The result is $\phi''(u)s(u)+2\phi'(u)s'(u)+\phi(u)s''(u)-\phi''(u)F'(u)-\phi'(u)F''(u)$. The value at zero of this second derivative is $\phi''(0)s(0)+2\phi'(0)s'(0) - \phi''(0)F'(0) - \phi'(0)F''(0) = 0$.

Differentiate a third time. This time we get $\phi'''(u)s(u) + 3\phi''(u)s'(u) + 3\phi'(u)s''(u)+\phi(u)s'''(u)-\phi'''(u)F'(u)-2\phi''(u)F''(u)-\phi'(u)F'''(u)$. The value at zero of this third derivative is $\phi'''(0)s(0)+3\phi''(0)s'(0)+3\phi'(0)s''(0)+\phi(0)s'''(0)-\phi'''(0)F'(0)-2\phi''(0)F''(0)-\phi'(0)F'''(0)$. This works out to be $-\phi''(0)F''(0)/2 < 0$.

Thus the function vanishes to second order at zero and its third derivative is negative. So to lowest order it looks like a positive constant time $-u^3$. This is enough to show that it is negative for small positive u.

5.8 The Hopf-Cole transformation for the Burgers equation

Recall that if we take the heat equation

$$\frac{\partial v}{\partial t} = \frac{1}{2}\sigma^2 \triangle v \tag{5.41}$$

and make the change of variables

$$v = \exp(-\frac{1}{m\sigma^2} w), \tag{5.42}$$

then we obtain the viscous quadratic Hamilton-Jacobi equation

$$\frac{\partial w}{\partial t} + \frac{1}{2m}|Dw|^2 = \frac{1}{2}\sigma^2 \triangle w. \tag{5.43}$$

This is called the Hopf-Cole transformation. We take the initial condition to be $w(y,t) = h(y)$. The corresponding initial condition for the heat equation is $\exp(-h(y)/(m\sigma^2))$.

We can also apply this to the gradient $u = Dw$. This gives the equation

$$\frac{\partial u}{\partial t} + \frac{1}{2m}Du^2 = \frac{1}{2}\sigma^2 \triangle u. \tag{5.44}$$

The initial condition is $u(y,t) = g(y) = Dh(y)$.

We can solve this equation by reducing it to the heat equation. The result is that

$$u(x,t) = -m\sigma^2 \frac{D_x v(x,t)}{v(x,t)} \tag{5.45}$$

where

$$v(x,t) = \int (2\pi\sigma^2 t)^{-\frac{n}{2}} \exp(-\frac{(x-y)^2}{2\sigma^2 t}) \exp(-\frac{h(y)}{m\sigma^2})\, dy. \tag{5.46}$$

This is the same as

$$u(x,t) = m \frac{\int \frac{x-y}{t} \exp(-\frac{K(x,y,t)}{\sigma^2})\, dy}{\int \exp(-\frac{K(x,y,t)}{\sigma^2})\, dy} \tag{5.47}$$

where

$$K(x,y,t) = \frac{(x-y)^2}{2t} + \frac{h(y)}{m}. \tag{5.48}$$

We may take the limit as σ^2 tends to zero. This is an application of Laplace's method. Fix x and t. Assume that there is only one point y for which $K(x,y,t) = (x-y)^2/(2t) + h(y)/m$ is minimal. Then by Laplace's method

$$u(x,t) \to m\frac{x-y}{t}. \tag{5.49}$$

as $\sigma^2 \to 0$. Thus the limiting solution is that u for which $x = y + \frac{1}{m}ut$. This is the Lax-Oleinik solution.

Next we can look at entropy $-\phi(u)$ for the Burger's equation. Again we take ϕ to be a convex function. We work in one dimension and take $m = 1$. The equation is

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \frac{1}{2}\sigma^2 \frac{\partial^2 u}{\partial x^2}. \tag{5.50}$$

Let $Y'(u) = \phi'(u)u$. Then

$$\frac{\partial \phi(u)}{\partial t} + \frac{\partial Y(u)}{\partial x} = \frac{1}{2}\sigma^2 \frac{\partial^2 \phi(u)}{\partial x^2} - \frac{1}{2}\sigma^2 \phi''(u)\left(\frac{\partial u}{\partial x}\right)^2. \tag{5.51}$$

It follows that

$$\frac{\partial \phi(u)}{\partial t} + \frac{\partial Y(u)}{\partial x} \leq \frac{1}{2}\sigma^2 \frac{\partial^2 \phi(u)}{\partial x^2}. \tag{5.52}$$

Now assume that we are considering solutions u that satisfy a uniform bound. Then the $\phi(u)$ also satisfy a uniform bound. This of course does not mean that the partial derivatives of $\phi(u)$ have to satisfy any kind of bound. However if we interpret the inequality in the sense of distributions, then the derivatives on the outside do not cause any problem. It follows that we can let σ^2 approach zero and conclude that

$$\frac{\partial \phi(u)}{\partial t} + \frac{\partial Y(u)}{\partial x} \leq 0 \tag{5.53}$$

in the sense of distributions. This says that the entropy is increasing. Furthermore, it shows the source of the entropy increase. It is the due to the fact that even when the viscosity σ^2 is small, the term $\sigma^2\phi''(u)(\partial u/\partial x)$ is large. In fact, the $(\partial u/\partial x)^2$ is going to be very large when the solution is approximating a shock.

Chapter 6

The Fourier transform

6.1 L^1 theory

In this chapter we consider spaces of functions defined on $\mathbf{R}^n$. The space C_c consists of continuous functions with compact support. The space C_0 consists of continuous functions that vanish at infinity. The space BC consists of all bounded continuous functions. The space L^∞ consists of all bounded measurable functions. In each case the norm is the supremum of the absolute value. It is easy to see that C_c is contained in C_0 which is contained in BC which is contained in L^∞. Furthermore, the closure of C_c in BC is C_0.

Even more important for our purposes are functions whose definition involves integration. The space L^1 consists of absolutely integrable functions. The norm on L^1 is the integral of the absolute value. The space L^2 consists of absolutely square integrable functions. The norm on L^2 is the square root of the integral of the square of the absolute value.

In the following we will think of two copies of $\mathbf{R}^n$. The first copy will be usually be regarded as associated with a space variable x. The measure used to compute integrals is the usual n dimensional Lebesgue measure dx. The other copy will usually be regarded as associated with a wave number variable k. The measure used to compute integrals is $dk/(2\pi)^n$. The space and wave number variables are regarded as dual, so that the dot product kx is dimensionless. Thus if space is measured in centimeters, the wave number is measured in inverse centimeters.

When $n = 1$ there is another common interpretation. The analog of the space variable is a time variable t, measured in seconds. The analog of the wave number variable is the angular frequency ω, measured in inverse seconds, or Hertz. Thus again ωt is dimensionless.

If f is in L^1 with respect to the space variable, then its Fourier transform is

$$\hat{f}(k) = \int e^{-ikx} f(x)\, dx. \tag{6.1}$$

The Fourier transform $\hat{f}$ is a function of the wave number variable. For f in

L^1, the function $\hat{f}$ is in L^∞, and we have $\|\hat{f}\|_\infty \leq \|f\|_1$. By the dominated convergence theorem $\hat{f}$ is actually in BC. The following result is even better. It us known as the Riemann-Lebesgue lemma.

Theorem 6.1 *If f is in L^1, then $\hat{f}$ is in C_0.*

Proof: Write

$$-\hat{f}(k) = \int e^{-ik(x+\pi/k)} f(x)\, dx = \int e^{-ikx} f(x - \pi/k)\, dx. \tag{6.2}$$

The second equality comes from a change of variables. Subtract this from the equation in the definition of the Fourier transform. This gives

$$2\hat{f}(k) = \int e^{-ikx}[f(x) - f(x - \pi/k)]\, dx. \tag{6.3}$$

Consequently,

$$2|\hat{f}(k)| \leq \int |f(x) - f(x - \pi/k)|\, dx. \tag{6.4}$$

Since translation is continuous in L^1, it follows that $\hat{f}(k)$ tends to zero as k goes to infinity.

If g is in L^1 with respect to the wave number variable, then its inverse Fourier transform is

$$\check{g}(x) = \int e^{ikx} g(k)\, \frac{dk}{(2\pi)^n}. \tag{6.5}$$

The inverse Fourier transform $\check{g}$ is a function of the space variable. The inverse Fourier transform has properties that correspond to the properties of the Fourier transform. The only difference is that we have to keep track of the sign in the exponent and the factors of 2π.

The key to proving properties of the Fourier transform is to first obtain a result for an approximate delta function. Let us take it to be

$$\delta_\epsilon(x) = (2\pi\epsilon^2)^{-\frac{n}{2}} \exp(-\frac{x^2}{2\epsilon^2}). \tag{6.6}$$

We already know that the integral is one. To compute the Fourier transform, note that the integral of the gradient $D_x[\exp(-ikx)\exp(-x^2/(2\epsilon^2)]$ is zero. However if we compute this gradient and integrate each term, we get that

$$D_k\hat{\delta}_\epsilon(k) = -\epsilon^2 k \hat{\delta}_\epsilon(k). \tag{6.7}$$

This equation can be solved, and the result is that

$$\hat{\delta}_\epsilon(k) = \exp(-\frac{\epsilon^2 k^2}{2}). \tag{6.8}$$

The same calculation, or a suitable change of variable, shows that the inverse Fourier transform of $\hat{\delta}_\epsilon(k)$ is the original $\delta_\epsilon(x)$. The fact that the inversion formula works in this case will turn out to imply that it works in general.

The following calculations will use convolution. It is not hard to show that the convolution of two L^1 functions is a function in L^1. Let h be an L^1 function. It is easy to compute directly that

$$(\delta_\epsilon * h)(x) = \int e^{-ikx} \hat{\delta}_\epsilon(k) \hat{h}(k) \, \frac{dk}{(2\pi)^n}. \tag{6.9}$$

The temptation is to let $\epsilon \to 0$ on both sides. However one must be careful; in general $\hat{h}$ will not be in L^1.

To see this, all one has to do is to consider a function h with a jump discontinuity. If $\hat{h}$ were in L^1, then h would have to be continuous, which is a contradiction. Thus to synthesize a function with discontinuities one needs to weigh high frequencies rather heavily.

Theorem 6.2 *Let h be an L^1 function. Then the inverse Fourier transform of $\hat{\delta}_\epsilon \hat{h}$ converges to h in the L^1 sense. In particular, h is determined by its Fourier transform.*

Proof: All we need to do is to check that $\delta_\epsilon * h$ converges to h in the L^1 sense. Compute

$$(\delta_\epsilon * h)(x) - h(x) = \int [h(x-y) - h(x)] \delta_\epsilon(y) \, dy = \int [h(x - \epsilon z) - h(x)] \delta_1(z) \, dz. \tag{6.10}$$

Integrate both sides with respect to x and interchange the order of integration. This shows that

$$\int |(\delta_\epsilon * h)(x) - h(x)| \, dx = \int \int |h(x - \epsilon z) - h(x)| \, dx \delta_1(z) \, dz. \tag{6.11}$$

The inner integral approaches zero by the continuity of translation in L^1. The outer integral thus approaches zero by the dominated convergence theorem.

It would be nice to have a condition that would guarantee that $\hat{f}$ is in L^1. This is given by the following lemma, which will be used later.

Lemma 6.1 *If h is in L^1 and in BC, and if $\hat{h} \geq 0$, then $\hat{h}$ is in L^1, and*

$$h(x) = \int e^{-ikx} \hat{h}(k) \, \frac{dk}{(2\pi)^n}. \tag{6.12}$$

Proof: We have

$$(\delta_\epsilon * h)(0) = \int \hat{\delta}_\epsilon(k) \hat{h}(k) \, \frac{dk}{(2\pi)^n}. \tag{6.13}$$

Let $\epsilon \to 0$. The left hand converges by the property of the approximate delta function, and the right hand converges by the monotone convergence theorem. This gives

$$h(0) = \int \hat{h}(k) \, \frac{dk}{(2\pi)^n}. \tag{6.14}$$

This shows that $\hat{h}$ is in L^1. From this we can use the dominated convergence theorem to get the result.

We now want to create a more flexible theory. First we need some machinery. Recall that the convolution $f * g$ of f and g in L^1 is defined by

$$(f * g)(x) = \int f(x - y)g(y)\, dy. \tag{6.15}$$

This is again an L^1 function. Furthermore, $\|f * g\|_1 \leq \|f\|_1 \|g\|_1$. It is easy to check that the Fourier transform of the convolution is the product of the Fourier transforms:

$$\widehat{(f * g)} = \hat{f}\,\hat{g}. \tag{6.16}$$

We also define the convolution adjoint of f to be

$$f^*(x) = \overline{f(-x)}. \tag{6.17}$$

The reason for the terminology is that convolution by f^* is the adjoint of convolution by f:

$$(f^* * g)(x) = \int \overline{f(y - x)} g(y)\, dy. \tag{6.18}$$

It is easy to check that the Fourier transform of the convolution adjoint of f is the complex conjugate of the Fourier transform of f.

6.2 The Plancherel theorem and L^2 theory

Now we want to assume that f and g are also in L^2. Then $f * g$ is in L^∞, by the Schwarz inequality. In fact, one can use the fact that translation is continuous in L^2 together with the Schwarz inequality to prove that $f * g$ is in BC.

The result of all this is that the Fourier transform of the continuous function $h = f^* * f$ is the positive function $\hat{h} = |\hat{f}|^2$. The lemma above thus proves the Plancherel theorem.

Theorem 6.3 *Let f be in L^1 and in L^2. Then*

$$\int |f(x)|^2\, dx = \int |\hat{f}(k)|^2 \frac{dk}{(2\pi)^n}. \tag{6.19}$$

Thus, the Fourier transform preserves the L^2 norm.

The Plancherel theorem allows us to define the Fourier transform of an arbitrary function f in L^2. Let f_m be a sequence of functions in L^1 and L^2 such that $f_m \to f$ in L^2 as $m \to \infty$. Thus, for instance, one could take f_m to be f inside a ball of radius m and 0 outside the ball. Then the f_m form a Cauchy sequence in L^2. By the Plancherel theorem the $\hat{f}_m$ form a Cauchy sequence in L^2. Since L^2 is a complete metric space, the $\hat{f}_m$ converge to a function $\hat{f}$ in L^2. This is the extended definition of the Fourier transform.

Define the inner product in L^2 of position space in the usual way as

$$\langle f, g\rangle = \int \overline{f(x)} g(x)\, dx. \tag{6.20}$$

The corresponding norm is $\|f\|_2 = \sqrt{\langle f, f\rangle}$. Define the inner product in L^2 of momentum space so that

$$\langle \hat{f}, \hat{g}\rangle = \int \overline{\hat{f}(k)} \hat{g}(k)\, \frac{dk}{(2\pi)^n}. \tag{6.21}$$

The corresponding norm is $\|\hat{f}\|_2 = \sqrt{\langle \hat{f}, \hat{f}\rangle}$.

It is an immediate consequence of the Plancherel theorem that $\|\hat{f}\|_2 = \|f\|_2$ for all f in L^2. That is, the Fourier transform preserves the L^2 norm. It is an isomorphism of Hilbert spaces.

Since in a complex vector space the norm determines the inner product, it follows that $\langle \hat{f}, \hat{g}\rangle = \langle f, g\rangle$ for all f and g in L^2. That is, the Fourier transform preserves the L^2 inner product.

All of this theory works just as well for the inverse Fourier transform. Thus we have the Fourier transform from L^2 of position space to L^2 of wave number space. We also have the inverse Fourier transform from L^2 of wave number space to L^2 of position space. All that remains to do is to prove that they are inverses of each other.

Theorem 6.4 *The Fourier transform on L^2 and the inverse Fourier transform on L^2 are inverses of each other.*

Proof: Let g be in L^1 and in L^2, and let h be in L^1 and in L^2. Then it is easy to check that $\langle \hat{g}, h\rangle = \langle g, \check{h}\rangle$. Since the inner product is continuous in L^2, it follows that for all g in L^2 and h in L^2 we have the same identity.

Take f in L^2 and let $h = \hat{f}$ in the above identity. Then we have $\langle g, f\rangle = \langle \hat{g}, \hat{f}\rangle = \langle g, \check{\hat{f}}\rangle$. Since g is arbitrary, we have $f = \check{\hat{f}}$. This proves that f is the inverse Fourier transform of the Fourier transform of f.

One can prove in a similar way that h is the Fourier transform of the inverse Fourier transform of h.

It may be worth summarizing exactly what is meant by these Fourier transforms. Let f be in L^2, and let f_M be f for $|x| \leq M$ and zero elsewhere. Then f_M is in L^1 and in L^2, and its Fourier transform $\hat{f}_M$ is absolutely convergent and represents a function in L^2. The definition of the Fourier transform $\hat{f}$ is that function such that $\|\hat{f} - \hat{f}_M\|_2 \to 0$ as $M \to \infty$.

Similarly, let $g = \hat{f}$ and g_N be g for $|k| \leq N$ and be zero elsewhere. Then g_N is in L^1 and in L^2, and its inverse Fourier transform $\check{g}_N$ is absolutely convergent and defines a function in L^2. The representation of f as a Fourier transform means that $\|f - \check{g}_N\|_2 \to 0$ as $N \to \infty$.

Example: Take dimension $n = 1$. Let $f(x) = 1$ for $|x| \leq a$ and $f(x) = 0$ otherwise. Then f is in both L^1 and L^2. Its Fourier transform $\hat{f}$ is given by

$$\hat{f}(k) = \int_{-a}^{a} \exp(-ikx)\, dx = 2\frac{\sin(ka)}{k}. \tag{6.22}$$

This is bounded and continuous and approaches zero at infinity. Furthermore, $\hat{f}$ is in L^2. However $\hat{f}$ is not in L^1. (It cannot be, since f is not continuous.) Therefore, the inverse Fourier transform is not absolutely convergent. All we can say is that

$$\int_{-\infty}^{\infty} |f(x) - \int_{-N}^{N} e^{ikx}\hat{f}(k)\,\frac{dk}{2\pi}|^2\, dx \to 0 \tag{6.23}$$

as $N \to \infty$.

On the other hand, the Plancherel theorem $\|f\|_2^2 = \|\hat{f}\|_2^2$ gives an equality between two absolutely convergent integrals. Explicitly, it says that

$$2a = \int_{-\infty}^{\infty} 4\frac{\sin^2(ka)}{k^2}\,\frac{dk}{2\pi}. \tag{6.24}$$

6.3 L^2 derivatives

Let f be in L^2. Denote the translate of f by a by f_a. Thus $f_a(x) = f(x-a)$. It is easy to verify that the Fourier transform $\hat{f}_a$ of the translate is given by multiplication by a phase:

$$\hat{f}_a(k) = \exp(-ika)\hat{f}(k). \tag{6.25}$$

This is the fundamental identity that underlies applications of Fourier transformations.

As an example, take the convolution of f with an L^1 function g. We have

$$(f * g)(x) = \int f(x-a)g(a)\, da. \tag{6.26}$$

The Fourier transform is thus

$$(\widehat{f * g})(k) = \int \exp(-ika)\hat{f}(k)g(a)\, da = \hat{f}(k)\hat{g}(k). \tag{6.27}$$

It is not difficult to show that if f is in L^2 and g is in L^1, then $f * g$ is in L^2. Furthermore, $\|f*g\|_2 \leq \|f\|_2\|g\|_1$. Correspondingly, $\hat{f}$ is in L^2, $\hat{g}$ is in L^∞ (even in C_0), and $\hat{f}\hat{g}$ is in L^2. And also $\|\hat{f}\hat{g}\|_2 \leq \|hatf\|_2\|\hat{g}\|_\infty$.

Let f be in L^2. Define the directional derivative $a \cdot Df$ of f in the L^2 sense to be the function in L^2 that is the limit of $[f(x+ta) - f(x)]/t$ in the L^2 sense, provided that the limit exists. By taking the Fourier transform, we see that

$$\widehat{a \cdot D}f(k) = ia \cdot k\hat{f}(k). \tag{6.28}$$

This result is so fundamental that it needs to be stated as a theorem.

Theorem 6.5 *Let f be in L^2. Then $a \cdot Df$ exists in the L^2 sense if and only if $ia \cdot k\hat{f}(k)$ is an L^2 function of k. In that case, $a \cdot Df$ is the inverse Fourier transform of $ia \cdot k\hat{f}(k)$.*

We can take the vectors a to be the unit basis vectors. Thus we conclude that the components of Df each exist in the L^2 sense if and only if the components of $ik\hat{f}(k)$ are each in L^2.

In view of this result, it seems only natural to define the Laplacian in the L^2 sense as follows. If f is in L^2, then $\triangle f$ is in L^2 if and only if $-k^2\hat{f}(k)$ is an L^2 function of k. In that case $\triangle f$ is the inverse Fourier transform of $-k^2\hat{f}(k)$.

6.4 The Poisson equation

The basic idea of solving a partial differential equation by the Fourier transform is to replace the partial differential operator by a polynomial in the wave number. Then solving the equation merely consists of dividing by the polynomial. The difficulty of course is in dealing with division by zero. For elliptic equations this problem is a nuisance but can sometimes be overcome, as we shall see in this section.

The safest equation from this point of view is the Poisson equation with a constant term $c > 0$. This is

$$\triangle u - cu + f = 0. \tag{6.29}$$

It represents equilibrium with a source f and a damping given by the term with $c > 0$. The damping should make the equilibrium automatic, and this is exactly what the Fourier transform gives.

Take f in L^2. The Fourier transform gives $(k^2 + c)\hat{u}(k) = \hat{f}(k)$. Thus

$$\hat{u}(k) = \frac{1}{k^2 + c}\hat{f}(k). \tag{6.30}$$

There is no division by zero. To solve the equation, all one has to do is to find the inverse Fourier transform $g(x)$ of $1/(k^2 + c)$. Then the solution is the convolution $g * f$.

This is easy to do explicitly in dimension $n = 1$. The Fourier transform of

$$g(x) = \frac{1}{2\sqrt{c}}e^{-\sqrt{c}|x|} \tag{6.31}$$

is easily computed, and turns out that $\hat{g}(k)$ is $1/(k^2 + c)$. Notice, however, that there is big trouble if we try to take c to be zero.

It is also not difficult to do a similar computation in dimension $n = 3$. The Fourier transform of

$$g(x) = \frac{1}{4\pi|x|}e^{-\sqrt{c}|x|} \tag{6.32}$$

may be computed by going to spherical polar coordinates. This reduces to a one dimensional integral which is much like the one dimensional case. The result is that for $n = 3$ we get the desired $\hat{g}(k)$ as $1/(k^2 + c)$. Notice that in this case taking the limit as c goes to zero is not so bad.

Let us look at the problem of the Poisson equation for all dimensions $n > 2$. The equation is

$$\triangle u + f = 0 \tag{6.33}$$

and represents equilibrium with a source, but with no damping.

Take f in L^2. The Fourier transform gives $k^2\hat{u}(k) = \hat{f}(k)$. Thus

$$\hat{u}(k) = \frac{1}{k^2}\hat{f}(k). \tag{6.34}$$

There is division by zero, but only at one point. If $n > 2$, then $1/k^2$ is locally integrable. So it looks like there is some hope for making sense of the Fourier transform.

A useful computational trick is to take $\epsilon > 0$ and write

$$\frac{1}{k^2}\exp(-\epsilon k^2) = \int_\epsilon^\infty \exp(-tk^2)\, dt. \tag{6.35}$$

This has two advantages. First, the function with $\epsilon > 0$ is in L^1. Second, one can compute its inverse Fourier transform.

The inverse Fourier transform is

$$\phi_\epsilon(x) = \int_\epsilon^\infty \frac{1}{(4\pi t)^{n/2}}\exp(\frac{x^2}{4t})\, dt = \frac{1}{4\pi^{\frac{n}{2}}}\frac{1}{|x|^{n-2}}\int_0^{\frac{x^2}{4\epsilon}} u^{\frac{n}{2}-2}\exp(-u)\, du. \tag{6.36}$$

This has the correct behavior for the solution of the Poisson equation for large x. However it is smoothed out for small x.

Now we can take the limit as ϵ approaches zero. This gives the fact that the solution of the Poisson equation is convolution by

$$\phi(x) = \frac{1}{4\pi^{\frac{n}{2}}}\Gamma(\frac{n}{2} - 2)\frac{1}{|x|^{n-2}}. \tag{6.37}$$

This checks with the previous result, since the reciprocal of the area of the unit sphere is

$$\frac{1}{n\alpha(n)} = \frac{\Gamma(\frac{n}{2} + 1)}{n\pi^{\frac{n}{2}}} = \frac{\Gamma(\frac{n}{2})}{2\pi^{\frac{n}{2}}} = \frac{(n-2)\Gamma(\frac{n}{2} - 1)}{4\pi^{\frac{n}{2}}}. \tag{6.38}$$

The preceding calculation works when $n > 2$. When $n = 1$ or $n = 2$ the function $1/k^2$ is not integrable near the origin. It is still possible to do a Fourier transform calculation, but this involves more delicate limiting operations, including infinite subtractions. The nice thing about dimensions $n > 2$ is that low frequencies (large distances) do not produce any trouble, even though one is dividing by zero at zero frequency.

6.5 The heat equation

Next we look at the heat equation, the standard example of a parabolic equation. We solve this by doing a Fourier transform only in the space variable. The partial differential equation is thus transformed into an ordinary differential equation in time for each wave number k. The equation is

$$\frac{\partial u}{\partial t} = \frac{1}{2}\sigma^2 \triangle u \tag{6.39}$$

with initial condition $u(x, 0) = h(x)$. The Fourier transform of this equation with respect to the space variable is

$$\frac{\partial \hat{u}(k,t)}{\partial t} = \frac{1}{2}\sigma^2(-k^2)\hat{u}(k,t) \tag{6.40}$$

with initial condition $\hat{u}(k,t) = \hat{h}(k)$. The solution of this equation may be computed for each fixed wave number k. This is the product

$$\hat{u}(k,t) = \exp(-\frac{1}{2}\sigma^2 k^2 t)\hat{h}(k). \tag{6.41}$$

If follows that the solution is the convolution

$$u(x,t) = (g_t * h)(x), \tag{6.42}$$

where

$$g_t(x) = (2\pi\sigma^2 t)^{-\frac{n}{2}} \exp(-\frac{x^2}{2\sigma^2 t}) \tag{6.43}$$

is the Gaussian. Note that the solution in wave number space has the high wave number damped out by the exponential factor. This causes the solution in position space to be smooth.

6.6 The Schrödinger equation

Next we look at the Schrödinger equation of quantum mechanics. This represents another category of partial differential equations, which we may call dispersive equations.

We solve this by doing a Fourier transform only in the space variable. The equation is

$$\frac{\partial u}{\partial t} = \frac{1}{2}i\sigma^2 \triangle u \tag{6.44}$$

with initial condition $u(x, 0) = h(x)$. The Fourier transform of this equation with respect to the space variable is

$$\frac{\partial \hat{u}(k,t)}{\partial t} = \frac{1}{2}i\sigma^2(-k^2)\hat{u}(k,t) \tag{6.45}$$

with initial condition $\hat{u}(k,t) = \hat{h}(k)$. The solution of this equation may be computed for each fixed wave number k. This is the product

$$\hat{u}(k,t) = \exp(-\frac{1}{2}i\sigma^2k^2t)\hat{h}(k). \tag{6.46}$$

If follows that the solution is the convolution

$$u(x,t) = (g_t * h)(x), \tag{6.47}$$

where

$$g_t(x) = (2\pi i\sigma^2 t)^{-\frac{n}{2}} \exp(i\frac{x^2}{2\sigma^2 t}) \tag{6.48}$$

is a complex Gaussian. The convolution integral makes sense when h is in L^1 as well as in L^2. Note that the solution in wave number space is not damped out at all by the complex exponential factor. There is no apparent smoothing, unless it were by some complicated method of oscillation.

The interpretation of the Schrödinger equation is that for each t the complex solution $u(x,t)$ is such that $|u(x,t)|^2$ is a probability density. This represents the probability of finding a particle at various regions of space.

The solution of the Schrödinger equation may be written in a form that exhibits its physical significance. We write the solution explicitly as

$$u(x,t) = (2\pi i\sigma^2 t)^{-\frac{n}{2}} \int \exp(i\frac{(x-y)^2}{2\sigma^2 t})h(y)\,dy. \tag{6.49}$$

We can expand the quadratic expression in the exponential and get three factors. Define three corresponding operators. First

$$M_t h(y) = \exp(i\frac{y^2}{2\sigma^2 t})h(y). \tag{6.50}$$

This is just a multiplication operator. It is a unitary operator from $L^2(dx)$ to itself. Second,

$$Ff(k) = \int \exp(-iky)f(y)\,dy \tag{6.51}$$

is the Fourier transform, defined by this formula on a dense subset of L^2. It may be interpreted as a unitary operator from $L^2(dx)$ to $L^2(dk/(2\pi)^n)$. Third,

$$S_t v(x) = (2\pi i\sigma^2 t)^{-\frac{n}{2}} \exp(i\frac{x^2}{2\sigma^2 t})v(\frac{x}{\sigma^2 t}). \tag{6.52}$$

This is a unitary operator from $L^2(dk/(2\pi)^n)$ to $L^2(dx)$.

Let $U_t h$ be the solution of the Schrödinger equation with initial condition h. Then the explicit solution we have found says that $U_t = S_t F M_t$.

Theorem 6.6 *Let $U_t h$ be the solution of the Schrödinger equation with initial condition h. Let $\hat{h}$ be the Fourier transform of h. Then the long term behavior of the solution is governed by the Fourier transform of the initial condition:*

$$\lim_{t\to\infty} \|U_t h - S_t \hat{h}\|_2 = 0. \tag{6.53}$$

Proof: Compute

$$\|U_t h - S_t \hat{h}\|_2 = \|S_t F M_t h - S_t F h\|_2. \tag{6.54}$$

This in turn is

$$\|S_t F M_t h - S_t F h\|_2 = \|S_t F(M_t h - h)\|_2 = \|(M_t h - h)\|_2 \tag{6.55}$$

since S_t and F are unitary. However the right hand side converges to zero as $t \to \infty$, by the dominated convergence theorem.

This theorem says that the solution of the Schrödinger equation satisfies

$$u(x,t) \sim (2\pi i \sigma^2 t)^{-\frac{n}{2}} \exp(i\frac{x^2}{2\sigma^2 t})\hat{h}(\frac{x}{\sigma^2 t}). \tag{6.56}$$

In particular, the probability density

$$|u(x,t)|^2 \sim (2\pi\sigma^2 t)^{-n}|\hat{h}(\frac{x}{\sigma^2 t})|^2. \tag{6.57}$$

The probability of finding a particle near x at time t depends on the value of the Fourier transform at wave number $k = x/(\sigma^2 t)$. Thus the velocity of the particles is related to the wave number by

$$\frac{x}{t} = \sigma^2 k \tag{6.58}$$

Different wave numbers travel with different velocities. This is why the equation is called dispersive.

In quantum mechanics the diffusion constant $\sigma^2 = \hbar/m$, where m is the mass of the particle and $\hbar$ is the rationalized version of Planck's constant. The last equation is often written as

$$p = \hbar k, \tag{6.59}$$

where $p = mx/t$ is the momentum.

6.7 Hyperbolic equations

Hyperbolic equations may also be solved by Fourier transformation in the space variable.

First we look at the transport equation

$$\frac{\partial u}{\partial t} + \mathbf{b} \cdot Du = cu \tag{6.60}$$

with $u(x,0) = g(x)$. The Fourier transform is

$$\frac{\partial \hat{u}(k,t)}{\partial t} + i\mathbf{b} \cdot k\, \hat{u}(k,t) = c\hat{u}(k,t) \tag{6.61}$$

with $\hat{u}(k, 0) = \hat{g}(k)$. The solution is

$$\hat{u}(k, t) = \exp(-i\mathbf{b} \cdot kt) \exp(ct)\hat{g}(k). \tag{6.62}$$

Since multiplication by a complex exponential function of wave number corresponds to a shift in space, the solution is

$$u(x, t) = h(x - \mathbf{b}t) \exp(ct). \tag{6.63}$$

Notice that the complex exponential does not damp high frequencies, and correspondingly there is no smoothing.

Next we look at the wave equation in one dimension. This is

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \tag{6.64}$$

with initial condition $u(x, 0) = g(x)$ and $\partial u/\partial t(x, 0) = h(x)$. The Fourier transform is

$$\frac{\partial^2 \hat{u}(k, t)}{\partial t^2} = -c^2 k^2 \hat{u}(k, t) \tag{6.65}$$

with initial condition $\hat{u}(x, 0) = g(x)$ and $\partial \hat{u}/\partial t(k, 0) = \hat{h}(k)$. The solution of the ordinary differential equation is

$$\hat{u}(k, t) = \cos(ckt)\hat{g}(k) + \frac{\sin(ckt)}{ck}\hat{h}(k). \tag{6.66}$$

Since the cosine is the sum of two complex exponentials, the first term gives a sum of two shifted functions. The second term may be computed by noting the that Fourier transform of an indicator function that is 1 on the interval from $-a$ to a and zero elsewhere is $2\sin(ka)/k$. Therefore the second term gives convolution by such a step function, with $a = ct$. We obtain the d'Alambert solution

$$u(x, t) = g(x - ct) + g(x + ct) + \int_{x-ct}^{x+ct} h(y)\, dy. \tag{6.67}$$

We can also look at the energy from the point of view of the Fourier transform. This is

$$E = \int \left[\left(\frac{\partial u}{\partial t}\right)^2 + c^2 \left(\frac{\partial u}{\partial x}\right)^2 \right] dx. \tag{6.68}$$

This can be computed to be

$$E = \int \left[|\hat{h}(k)|^2 + c^2 k^2 |\hat{g}(k)|^2 \right] \frac{dk}{2\pi}. \tag{6.69}$$

This shows not only that energy is conserved, but the energy for each frequency interval is conserved.

6.8 L^2 Sobolev inequalities

This section uses the L^2 theory of the Fourier transform to prove some elementary Sobolev inequalities. These show that the condition that a function in L^2 has L^2 partial derivatives has implications for the regularity of the function, even though the partial derivatives may not exist in the classical pointwise sense. One interesting feature is that the results depend on the dimension of space.

Theorem 6.7 *Suppose that $D^\alpha f$ is in $L^2(\mathbf{R}^n)$ for all derivatives of order $|\alpha| \leq s$. Suppose also that $s > n/2$. Then $\hat{f}$ is in $L^1(\mathbf{R}^n)$, and hence f is in $C_0(\mathbf{R}^n)$. Furthermore, the maximum of $|f|$ is bounded by the L^1 norm of $\hat{f}$, and this in turn is bounded by a constant times an expression involving the L^2 norms of the derivatives $D_\alpha f$ for $|\alpha| \leq s$.*

Proof: The fact that $D^\alpha f$ is in L^2 for all $|\alpha| \leq s$ is equivalent to the fact that $k^\alpha \hat{f}(k)$ for all $|\alpha| \leq s$. This in turn is equivalent to the fact that $(1 + k^2)^{\frac{s}{2}} \hat{f}(k)$ is in L^2.

The essential step is to write

$$\hat{f}(k) = \frac{1}{(1 + k^2)^{\frac{s}{2}}} \cdot (1 + k^2)^{\frac{s}{2}} \hat{f}(k). \tag{6.70}$$

We have just seen that the second factor is in L^2. Since $s > n/2$, the first factor is also L^2. Since the product of two L^2 functions is in L^1, it follows that $\hat{f}(k)$ is in L^1. By the inversion formula and the Riemann Lebesgue lemma f is in C_0.

The theorem just proved is important for non-linear partial differential equations. The reason is that one starts with an L^2 condition on derivatives, which one gets from energy conditions. However the result is an L^∞ estimate, a bound on the maximum size of the function. The nice feature of L^∞ estimates is that they are preserved under non-linear operations. For instance, look at the operation of replacing f by $\phi(f)$, where f is a continuous function. If we have a class of functions f for which $\|f\|_\infty \leq M$ is bounded, and if ϕ is a continuous function, then the norms $\|\phi(f)\|_\infty \leq \max_{|z| \leq M} |\phi(z)|$ are also bounded.

There is also a corresponding result for higher derivatives.

Corollary 6.1 *Suppose that D^α is in $L^2(\mathbf{R}^n)$ for all derivatives of order $|\alpha| \leq s + r$. Suppose also that $s > n/2$. Then $D^\beta f$ is in $C_0(\mathbf{R}^n)$ for all derivatives of order $|\beta| \leq r$. In particular f has continuous partial derivatives of order r.*

Example: Consider the heat equation. We have seen that its solution in the Fourier transform representation is

$$\hat{u}(k, t) = \exp(-\frac{1}{2}\sigma^2 k^2 t)\hat{h}(k). \tag{6.71}$$

Consider $t > 0$. By the Schwarz inequality, if $\hat{h}$ is in L^2, then

$$(1 + k^2)^{\frac{m}{2}} \hat{u}(k, t) = (1 + k^2)^{\frac{m}{2}} \exp(-\frac{1}{2}\sigma^2 k^2 t)\hat{h}(k) \tag{6.72}$$

is in L^1 for each m. This shows that the solution is C^∞ in the space variables. If we take any number of time derivatives, this pulls down the corresponding number of factors of k^2, but the result is still in L^1. This is enough to show that for $t > 0$ the solution

$$u(x,t) = \int \exp(ikx)\exp(-\frac{1}{2}\sigma^2 k^2 t)\hat{h}(k)\,\frac{dk}{(2\pi)^n} \tag{6.73}$$

is smooth in both space and time.

We can use these ideas to define new Hilbert spaces. The Sobolev space $H^s(\mathbf{R}^n)$ is defined for $s \geq 0$ as the Hilbert space of all f in $L^2(\mathbf{R}^n)$ such that $(1+k^2)^{\frac{s}{2}}\hat{f}$ is in L^2. Its norm is given by

$$\|f\|^2_{H^s} = \|(1-\triangle)^{\frac{s}{2}}f\|_2 = \int (1+k^2)^s |\hat{f}(k)|^2\,\frac{dk}{(2\pi)^n}. \tag{6.74}$$

The result we have just proved is that if $s > n/2$, then the Sobolev space $H^s(\mathbf{R}^n)$ is contained in $C_0(\mathbf{R}^n)$, and the inclusion is continuous.

Later on we shall need to deal with Sobolev spaces where there are relatively few derivatives, and where consequently the functions are not particularly smooth. On important special case is $s = 1$. Then the norm is

$$\|f\|^2_{H^1} = \|f\|_2^2 + \|Df\|_2^2 = \int (1+k^2)|\hat{f}(k)|^2\,\frac{dk}{(2\pi)^n}. \tag{6.75}$$

This expression resembles expressions for energy often encountered in the context of partial differential equations.

Functions in $H^1(\mathbf{R}^n)$ need not be particularly smooth when n is 2 or more. We can see this most easily when $n > 2$. In that case, a spike function u with a local singularity like $1/r^a$ with $a > 0$ can have Du in L^2, even though it is discontinuous at the origin. The reason is that the $D(1/r^a) = -a/r^{a+1}\,x/r$. The square of the length of this vector is a constant times $1/r^{2a+2}$. When this is integrated in polar coordinates with respect to the $r^{n-1}\,dr$ measure, the result is convergent near the origin, provided that $2a < n-2$.

It might seem that one point is not so bad. However we enumerate a dense set of points x_k in $\mathbf{R}^n$, $k = 1,2,3,\ldots$. We can then translate each spike to an origin centered at x_k, getting spike functions u_k. We can then multiply u_k by $1/2^k$ and sum. This gives a function that is discontinuous everywhere, yet is in the Hilbert space $H^2(\mathbf{R}^n)$ for $n > 2$.

Another important special case is $s = 2$. In this case.

$$\|f\|^2_{H^2} = \|(1-\triangle)f\|_2^2 = \int (1+k^2)^2|\hat{f}(k)|^2\,\frac{dk}{(2\pi)^n}. \tag{6.76}$$

This Sobolev space is the natural domain for the Laplace operator, if we want the result of applying the operator to be in L^2. These functions are somewhat smoother. Thus when $n = 3$ they are even bounded and continuous.

These Sobolev spaces will be heavily used in connection with energy methods for solving partial differential equations.

Chapter 7

The Dirichlet Laplacian

7.1 Sobolev spaces for open subsets

Let U be an open set. Let $C_c^\infty(U)$ be the space of all smooth functions with compact support within U. For $s \geq 0$, let $H_0^s(U)$ be the closure of $C_c^\infty(U)$ in the Sobolev space H^s. Thus $H_0^s(U)$ consists of functions in $L^2(U)$ with partial derivatives up to order s belonging to $L^2(U)$. Furthermore, the functions in $H_0^s(U)$ all vanish at the boundary of U, in the sense that they are approximated by functions that vanish near the boundary of U. Each such function can thus be extended to all of $\mathbf{R}^n$, so that the usual tools of Fourier analysis continue to apply.

The norm on the Sobolev space H^1 is $\|(1-\triangle)^{\frac{1}{2}}u\|_2 = \sqrt{\|u\|_2 + \|Du\|_2}$. The same norm works for each of the spaces $H_0^1(U)$, where U is an open set.

7.2 The Laplacian with dissipation

In this section we show that if $\gamma > 0$, then the equation

$$\triangle u - \gamma u + f = 0 \tag{7.1}$$

in U with Dirichlet boundary conditions on the boundary of U always has a unique solution. This is the equation for equilibrium with a source, when the boundary condition is zero. The condition that $\gamma > 0$ is a condition that ensures a certain amount of dissipation. Thus it is not surprising that there is an equilibrium in this case. We shall see in a later section that if the region U has finite measure, then there is an equilibrium even without dissipation.

The solution that will be constructed is a kind of weak solution. The proof that it is a solution in a more classical sense involves more technicalities. These will not be treated in this chapter.

The idea is to use the Hilbert space norm whose square is

$$\|u\|_{H_0^2}^2 = \|Du\|_2^2 + \gamma\|u\|_2. \tag{7.2}$$

This is a kind of energy norm. The Dirichlet principle states that the solution should be obtained by minimizing the energy

$$E(u) = \frac{1}{2}\|u\|^2_{H^1_0} - \langle f, u\rangle, \tag{7.3}$$

where the inner product on the second term is the usual L^2 inner product.

If there is a minimal energy function, then it should satisfy the equation

$$\langle u, h\rangle_{H^1_0} - \langle f, h\rangle = 0 \tag{7.4}$$

for all h in the Sobolev space. This says that

$$\langle Du, Dh\rangle + \gamma\langle u, h\rangle - \langle f, h\rangle = 0 \tag{7.5}$$

for all h in the Sobolev space. This is what we will mean by a weak solution of the problem.

Theorem 7.1 *Let U be an open subset of $\mathbf{R}^n$. Suppose that $\gamma > 0$. Then the equation*

$$\triangle u - \gamma u + f = 0 \tag{7.6}$$

in U with Dirichlet boundary conditions on the boundary of U always has a unique weak solution.

Proof: Let $L(h) = \langle f, h\rangle$. This is clearly a continuous linear functional on the Hilbert space $L^2(U)$, by the Schwarz inequality. Since $\gamma > 0$, it is also a continuous linear functional on the Hilbert space $H^1_0(U)$. Therefore, by the Riesz representation theorem, there is an element u in $H^1_0(U)$ that represents the linear functional. That is,

$$\langle u, h\rangle_{H^1_0} = L(h) \tag{7.7}$$

for all h in $H^1_0(U)$. This is the desired weak solution.

It would seem that this proof has nothing to do with Dirichlet's principle. However in the next section we shall see that it is really the same thing.

7.3 Dirichlet's principle and the Riesz representation theorem

In this section we give a natural proof of the Riesz representation theorem for continuous linear functionals on Hilbert space. This proof uses the idea of Dirichlet's principle, so it is makes the connection outlined in the previous section.

Theorem 7.2 *Let L be a continuous linear functional on a Hilbert space $\mathcal{H}$. Then there is an element u of $\mathcal{H}$ such that*

$$L(h) = \langle u, h\rangle \tag{7.8}$$

for all h in $\mathcal{H}$. In other words, the continuous linear functional is represented by the vector u by use of the inner product.

Proof: Consider the energy

$$E(u) = \frac{1}{2}\|u\|^2 - L(u). \tag{7.9}$$

Since L is continuous, it satisfies a bound $|L(u)| \leq C\|u\|$. It follows easily that the energy is bounded below by $-1/2C^2$. Let M be the greatest lower bound on the energy. Then there exists a sequence u_n such that $E(u_n) \to M$ as $n \to \infty$. If we could show that this sequence converges to some u in the Hilbert space, then by continuity $E(u) = M$, and we have found a minimum point.

To do this, we use the parallelogram identity

$$\frac{1}{4}\|u_m - u_n\|^2 = E(u_m) + E(u_n) - 2E((u_m + u_n)/2). \tag{7.10}$$

(If we think of the right hand side as a kind of second difference approximation, then this says that the energy is convex in a very strong sense.) It follows that

$$\frac{1}{4}\|u_m - u_n\|^2 \leq E(u_m) + E(u_n) - 2M. \tag{7.11}$$

If we take m and n large enough, then the right hand side of this inequality gets close to zero. This is enough to show that the sequence u_n is a Cauchy sequence. Since a Hilbert space is a complete metric space, the sequence must converge to some u in the space.

Since we have a minumum point with $E(u) = M$, it follows that $E(u+h) \geq M$ for all vectors h in the Hilbert space. This says that

$$\frac{1}{2}\|u\|^2 + \langle u, h\rangle + \frac{1}{2}\|h\|^2 - L(u) - L(h) \geq M. \tag{7.12}$$

Thus

$$\langle u, h\rangle + \frac{1}{2}\|h\|^2 - L(h) \geq 0. \tag{7.13}$$

Take $t > 0$ and replace h by th in this inequality. Divide by t and take the limit ad t approaches zero. This shows that

$$\langle u, h\rangle - L(h) \geq 0. \tag{7.14}$$

Replace h by $-h$ in the inequality. This shows that also

$$\langle u, h\rangle - L(h) \leq 0. \tag{7.15}$$

This completes the proof.

The ultimate power of this result in applications comes from the deep fact that L^2 is a Hilbert space. This uses essentially the theory of Lebesgue integration.

7.4 Finite elements

In numerical computation one wants to approximate the solution that is the minimum of the energy functional $E(u)$ defined for u in the Sobolev space $H_0^1(U)$. The idea is to take a finite dimensional subspace M of the Sobolev space that is large enough to provide good approximations to arbitrary functions in the Sobolev space. Then one minimizes $E(u)$ for u in M. The minimum value will in general be larger than the minimum on the Sobolev space, but it is an upper bound. Furthermore, the function u in M may turn out to be a good approximation to the solution of the original problem.

Here is one method that one could accomplish this. For simplicity we describe it in the case $n = 2$. Say that $\bar{U}$ is a union of finitely many triangular regions (elements) that overlap only at their boundaries. Let I be the set of all vertices (nodes) of these triangles that belong to the interior U. If p is a vertex in I, define the function ϕ_p to satisfy $\phi_p(p) = 1$, $\phi_p(q) = 0$ for all $q \leq p$, and ϕ_p linear on each triangle. Thus ϕ_p is non-zero only on those triangular regions that have a vertex at p. It is a function whose graph is a pyramid. Each function ϕ_p is continuous, vanishes at the boundary, and has a weak derivative in L^2. Thus it belongs to the Sobolev space $H_0^1(U)$. The finite dimensional subspace M is taken to be the subspace spanned by these basis vectors. Thus it consists of continuous piecewise linear functions that vanish at the boundary.

Such a basis is convenient because if vertices do not belong to the same triangle, then the corresponding basis vectors are orthogonal in the Sobolev space. It follows that the matrix of the quadratic form has many zeros.

In practice one may wish to use piecewise polynomial functions instead of piecewise linear functions. This is more complicated, but it may give more accurate results for the same computational effort.

7.5 Equivalent norms on the Sobolev space

The norm on the Sobolev space H_0^1 up to now has been defined to be given by $\|(1-\triangle)^{\frac{1}{2}}u\|_2 = \sqrt{\|u\|_2 + \|Du\|_2}$. The next result shows that when the measure of the open set is finite, then a more convenient choice of norm is possible. This result is one version of the Poincaré inequality. Notice that it is essential that we impose zero boundary conditions; otherwise a constant function would violate the inequality.

Theorem 7.3 *If U is an open subset of $\mathbf{R}^n$ with finite measure, then there exists a constant C depending on n and on the measure of U such that for all u in $H_0^1(U)$ we have*

$$\|u\|_2 \leq \|Du\|_2. \tag{7.16}$$

Proof for the case $n > 2$: Let χ be the indicator function of U. Since u is in L^2 and vanishes outside of U, it follows from the Schwarz inequality that u is in L^1 with norm bounded by $\|\chi\|_2\|u\|_2$. Hence the Fourier transform $\hat{u}$ is in

L^∞ with

$$\|\hat{u}\|_\infty \leq \|u\|_1 \leq \|\chi\|_2 \|u\|_2. \tag{7.17}$$

Let ϕ be the indicator function of the set $|k| \leq 1$. Then we have

$$\|u\|_2^2 = \|\hat{u}(k)\|_2^2 = \|\phi(k)\hat{u}(k)\|_2^2 + \|(1-\phi(k))\hat{u}(k)\|^2. \tag{7.18}$$

The first term is estimated by

$$\|\phi(k)\hat{u}(k)\|_2^2 \leq \|\hat{u}(k)\|_\infty \|\phi(k)\hat{u}\|_1 \leq \|u(k)\|_\infty \|\frac{1}{|k|}\phi(k)\|_2 \||k|\hat{u}(k)\|_2. \tag{7.19}$$

The second term is obviously bounded by

$$\|(1-\phi(k))\hat{u}(k)\|_2^2 \leq \|\hat{u}(k)\|_2 \||k|\hat{u}(k)\|_2. \tag{7.20}$$

If we put these together, we see that

$$\|u\|_2^2 \leq \|u\|_2 \left(\|\chi\|_2 \|\frac{1}{k}\|_2 + 1 \right) \|Du\|_2. \tag{7.21}$$

This Fourier transform argument does not work for $n \leq 2$, since the L^2 norm of $1/|k|$ is infinite in these cases. However the result that $\|u\|_2 \leq C\|Du\|_2$ for functions with compact support within a bounded open set is still true.

This is elementary for the case $n = 1$, by the following lemma.

Lemma 7.1 *Consider the case of dimension $n = 1$. If u has compact support and u' is in L^1, then u is in L^∞.*

Proof: This is the fundamental theorem of calculus. We have

$$u(x) = \int_{-\infty}^{x} u'(u)\, dy \tag{7.22}$$

and

$$u(x) = -\int_{x}^{\infty} u'(u)\, dy. \tag{7.23}$$

It follows immediately that

$$|u(x)| \leq \frac{1}{2} \int_{-\infty}^{\infty} |u'(y)|\, dy. \tag{7.24}$$

The lemma is applied by noting that if u' is in L^2 on a bounded interval, then by the Schwarz inequality u' is in L^1. Hence by the lemma u is in L^∞. Since the interval is bounded, u is also in L^2. Tracing through the inequalities, we see that

$$\|u\|_2 \leq \frac{1}{2} \operatorname{meas}(U) \|u'\|_2. \tag{7.25}$$

For $n = 2$ the result is a consequence of the Sobolev inequalities of the next chapter. Here is a quick proof for the case of dimension $n = 2$. First we need a lemma.

Lemma 7.2 *Consider the case of dimension $n = 2$. If u has compact support and Du is in L^1, then u is in L^2.*

Proof: This follows from the fundamental theorem of calculus applied to each of the two variables separately. For each x_1 we have

$$|u(x_1, x_2)| \leq \frac{1}{2} \int |Du(x_1, x_2)| \, dx_2 \tag{7.26}$$

and for each x_2 we have

$$|u(x_1, x_2)| \leq \frac{1}{2} \int |Du(x_1, x_2)| \, dx_1 \tag{7.27}$$

Multiply these inequalities and integrate with respect to both variables. This proves that $\|u\|_2 \leq \frac{1}{2}\|Du\|_1$.

In order to apply the lemma, observe that if U is an open set of finite measure, and if Du is in $L^2(U)$, then Du is in $L^1(U)$. It follows from the lemma that u is in $L^2(U)$. Furthermore, for $n = 2$

$$\|u\|_2 \leq \frac{1}{2}(\text{meas}(U))^{\frac{1}{2}} \|Du\|_2. \tag{7.28}$$

In any case, the significance of the theorem is the following. When U has finite measure, then the Sobolev norm $\|(1 - \triangle)^{\frac{1}{2}})u\|_2 = \sqrt{|u|_2^2 + \|Du\|_2}$ on $H_0^1(U)$ may be replaced by the norm $\|(-\triangle u)^{\frac{1}{2}} u\|_2 = \|Du\|_2$ only involving the derivative. This is an equivalent norm; the notion of convergence is not affected. Thus the norm $\|Du\|_2$ is a Hilbert space norm.

7.6 The Laplacian without dissipation

In this section we show that if U has finite measure, then the equation

$$\triangle u + f = 0 \tag{7.29}$$

in U with Dirichlet boundary conditions on the boundary of U always has a unique solution. This is the equation for equilibrium with a source, when the boundary condition is zero. The reason for equilibrium in this case is that the boundary is near enough to provides an environment that stabilizes the sytem

The idea is to use the Hilbert space norm $\|Du\|_2$. We have seen that this is equivalent to the Sobolev norm, and hence it is the norm on a Hilbert space. The Dirichlet principle states that the solution should be obtained by minimizing the energy

$$E(u) = \frac{1}{2}\|Du\|_2^2 - \langle f, u \rangle, \tag{7.30}$$

where the inner product on the second term is the usual L^2 inner product.

If there is a minimal energy function, then it should satisfy the equation

$$\langle Du, Dh \rangle - \langle f, h \rangle = 0 \tag{7.31}$$

for all h in the Sobolev space. This says that

$$\langle Du, Dh\rangle - \langle f, h\rangle = 0 \tag{7.32}$$

for all h in the Sobolev space. This is what we will mean by a weak solution of the problem.

Theorem 7.4 *Let U be an open subset of $\mathbf{R}^n$. Suppose that U has finite measure. Then the equation*

$$\triangle u + f = 0 \tag{7.33}$$

in U with Dirichlet boundary conditions on the boundary of U always has a unique weak solution.

Proof: Let $L(h) = \langle f, h\rangle$. This is clearly a continuous linear functional on the Hilbert space $L^2(U)$, by the Schwarz inequality. Since the measure of U is finite, it is also continuous linear functional on the Sobolev space with norm $\|Du\|_2$. Therefore, by the Riesz representation theorem, there is an element u in the Sobolev space that represents the linear functional. That is,

$$\langle Du, Dh\rangle = L(h) \tag{7.34}$$

for all h in the Sobolev space. This is the desired weak solution.

7.7 Positive elliptic operators

Essentially the same proof works for proving the existence and uniqueness of solutions of a much more general class of equations. Let A be a bounded function whose values $A(x)$ are positive definite n by n matrices. Assume that the eigenvalues of all the $A(x)$ are bounded below by a constant $\theta > 0$. Let c be a bounded real function such that $c(x) \geq 0$ for each x. Consider the equation

$$\operatorname{div} \mathbf{J} - c(x)u = f, \tag{7.35}$$

where

$$\mathbf{J} = A(x)Du. \tag{7.36}$$

This says that the current can depend on the gradient of the solution in a complicated way. Furthermore, the dissipation $-c(x)$ can take a complicated form. The restriction that $A(x)$ be symmetric seems arbitrary, but Onsager showed that this is related to general considerations of time reversal.

With this more general setting, one can consider the Dirichlet problem in the Sobolev space space $H_0^1(U)$, where U is a set of finite measure. The energy is now

$$E(u) = \frac{1}{2}\int Du(x)\cdot A(x)Du(x) + \frac{1}{2}\int c(x)|u(x)|^2 - \int f(x)u(x)\,dx. \tag{7.37}$$

The hypotheses ensure that this gives an equivalent norm on the Sobolev space. So again the Dirichlet principle (in the form of the Riesz representation theorem) applies to give a unique weak solution.

Chapter 8

Sobolev inequalities

8.1 L^p spaces

The basic Jensen's inequality says a convex function of an average is less than or equal to the average of the convex function.

We write $E[Y]$ for the expectation (mean, average) of Y. The fundamental properties of expectation are that it is linear, order-preserving ($Y \leq Z$ implies $E[Y] \leq E[Z]$), and that $E[c] = c$.

Theorem 8.1 *If ϕ is a smooth convex function, then*

$$\phi(E[X]) \leq E[\phi(X)]. \tag{8.1}$$

Proof: Since ϕ is convex, for each a we have $\phi(a) + \phi'(a)(x-a) \leq \phi(x)$. Thus $\phi(E[X]) + \phi'(E[X])(X - E[X]) \leq \phi(X)$. Take expectations of both sides of the inequality.

The most important special case is when $\phi(x) = e^x$. In that case this is called the inequality of the geometric and arithmetic mean. Explicitly:

$$e^{E[X]} \leq E[e^X]. \tag{8.2}$$

Alternatively, for $Y > 0$ we have

$$e^{E[\log(Y)]} \leq E[Y]. \tag{8.3}$$

The left hand side is the geometric mean; the right hand side is the arithmetic mean.

For each p with $1 \leq p \leq \infty$ there is a conjugate exponent q with

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{8.4}$$

One fundamental convexity inequality is that for each conjugate pair p and q we have for $a \geq 0$ and $b \geq 0$ the estimate

$$ab \leq \frac{1}{p}a^p + \frac{1}{q}b^q. \tag{8.5}$$

This follows immediately from the inequality of the geometric and arithmetic mean. Let $X = s$ with probability $1/p$ and $X = t$ with probability $1/q$. Then

$$e^{\frac{1}{p}s+\frac{1}{q}t} \leq \frac{1}{p}e^s + \frac{1}{q}e^t. \tag{8.6}$$

Take $e^s = a^p$ and $e^t = b^q$.

Remark: One must take some care with the right hand side if p or q is infinite. Consider, for example the case when $p = \infty$. If $a \leq 1$, then $(1/p)a^p = 0$. On the other hand, if $a > 1$, then $(1/p)a^p = \infty$. Then the inequality is just giving the obvious information that if $a \leq 1$, then $ab \leq b$.

For each p with $1 \leq p < \infty$ we want to consider the Banach spaces L^p of functions with norm

$$\|f\|_p = \left(\int |f(x)|^p \, dx\right)^{\frac{1}{p}} < \infty. \tag{8.7}$$

When $p = \infty$ we take L^∞ to be the space of essentially bounded functions, and the norm $\|f\|_\infty$ is the essential bound. That is, $\|f\|_\infty$ is the least M such that $|f(x)| \leq M$ for almost every x.

We want to argue that a finite L^p norm for large p implies good local properties of a function. The first evidence in this direction is the following Chebyshev inequality. It says that a small L^p norm implies that the function cannot be big on too large a set.

Theorem 8.2 *Let f be in L^p. For each a with $0 \leq a < \infty$ let A_a be the set of all x with $|f(x)| \geq a$. Then*

$$a^p \operatorname{meas}(A_a) \leq \|f\|_p. \tag{8.8}$$

This idea may be used to justify the notation for the case $p = \infty$.

Theorem 8.3 *Suppose that f is in L^q for some $q < \infty$. Then*

$$\lim_{p\to\infty} \|f\|_p = \|f\|_\infty. \tag{8.9}$$

Proof: We need inequalities in two different directions. To bound the L^∞ norm in terms of the L^p norm, we let $a < \|f\|_\infty$. Let A_a be the set of x for which $|f(x)| \geq a$. Then by definition of the L^∞ norm, $\operatorname{meas}(A_a) > 0$. Also, $a^p \operatorname{meas}(A_a) \leq \|f\|_p^p$, so

$$a(\operatorname{meas} A_a)^{\frac{1}{p}} \leq \|f\|_p. \tag{8.10}$$

To bound the L^p norm in terms of the L^∞ norm, we need to use the hypothesis that f is in L^q for some $q < \infty$. Then for $p \geq q$ we have $|f(x)|^p \leq |f(x)|^q \|f\|_\infty^{p-q}$ for almost every x. Hence we get the second inequality

$$\|f\|_p \leq \|f\|_q^{\frac{q}{p}} \|f\|_\infty^{1-\frac{q}{p}}. \tag{8.11}$$

The two inequalities together show that for large p the norm $\|f\|_p$ is squeezed between a and $\|f\|_\infty$. Since a is an arbitrary number less than $\|f\|_\infty$, this proves the result.

The following theorem is even more fundamental. It says that for a finite measure space the L^p spaces get smaller as p increases. Thus for finite measure space the condition of being in L^p is stronger when p is larger. A function in L^∞ is automatically in L^2, and a function in L^2 is automatically in L^1. In particular, this applies when the measure space is a bounded region with Lebesgue measure. So this result gives the local comparison of L^p spaces.

Theorem 8.4 *Suppose that the measure of the space U is finite. Then if $r \leq p$ and f is in L^p, then also f is in L^r. Furthermore,*

$$\|f\|_r \leq \mathrm{meas}(U)^{\frac{1}{q}} \|f\|_p, \tag{8.12}$$

where $1/p + 1/q = 1/r$.

Proof: This is a nice application of Jensen's inequality. Define the expectation of g to be

$$E[g] = \frac{1}{\mathrm{meas}(U)} \int_U g(x)\, dx. \tag{8.13}$$

Take $\phi(t) = t^{\frac{p}{r}}$. This is a convex function. Hence by Jensen's inequality

$$\left(\frac{1}{\mathrm{meas}(U)} \int |f(x)|^r\, dx\right)^{\frac{p}{r}} \leq \frac{1}{\mathrm{meas}(U)} \int |f(x)|^p\, dx. \tag{8.14}$$

This is equivalent to the statement of the theorem.

Consider two functions f and g with $\|f\|_p = 1$ and $\|g\|_q = 1$. Then by the geometric-arithmetic mean inequality

$$\int |f(x)g(x)|\, dx \leq \int \frac{1}{p}|f(x)|^p\, dx + \int \frac{1}{q}|g(x)|^q\, dx = \frac{1}{p} + \frac{1}{q} = 1. \tag{8.15}$$

This estimate leads immediately to the most basic form of Hölder's inequality.

Theorem 8.5 *If p and q are conjugate exponents with*

$$\frac{1}{p} + \frac{1}{q} = 1, \tag{8.16}$$

and if f is in L^p and g is in L^q, then fg is in L^1 with

$$\|fg\|_1 \leq \|f\|_p \|g\|_q. \tag{8.17}$$

Remark: How can one remember such relations between conjugate exponents as in the Hölder inequality? Dimensional analysis can be helpful. Think of f and g as dimensionless. However dx has dimension of length to the nth power.

Thus each integral has dimension of length to the nth power. For the dimensions to coincide on both sides, one must have

$$n = \frac{n}{p} + \frac{n}{q}. \tag{8.18}$$

The following corollary is convenient when one has the product of two functions f in L^p and g in L^q, but these are not necessarily conjugate exponents. Then one can compute an r such that the product fg is in L^r. The inequality makes sense even when r turns out to be less than one. However we need $r \geq 1$ if we are going to use the fact that $\|h\|_r$ is a norm.

Corollary 8.1 *If $r \leq p \leq \infty$ and $r \leq q \leq \infty$ and*

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{r}, \tag{8.19}$$

and if f is in L^p and g is in L^q, then fg is in L^r with

$$\|fg\|_r \leq \|f\|_p \|g\|_q. \tag{8.20}$$

It is pleasant to note that all this generalizes to several factors. If

$$\frac{1}{p_1} + \cdots + \frac{1}{p_m} = 1, \tag{8.21}$$

then the geometric-arithmetic mean inequality gives

$$a_1 \cdots a_m \leq \frac{1}{p_1} a_1^{p_1} + \cdots + \frac{1}{p_m} a_m^{p_m}. \tag{8.22}$$

This leads to the generalized Hölder inequality and its corollary.

Theorem 8.6 *Suppose*

$$\frac{1}{p_1} + \cdots + \frac{1}{p_m} = 1. \tag{8.23}$$

Then

$$\|f_1 \cdots f_m\|_1 \leq \|f_1\|_{p_1} \cdots \|f_m\|_{p_m}. \tag{8.24}$$

Corollary 8.2 *Suppose*

$$\frac{1}{p_1} + \cdots + \frac{1}{p_m} = \frac{1}{r}. \tag{8.25}$$

Then

$$\|f_1 \cdots f_m\|_r \leq \|f_1\|_{p_1} \cdots \|f_m\|_{p_m}. \tag{8.26}$$

8.2 Convolution and L^p

It is also helpful to have information on convolution of functions in L^p. Recall that the convolution of g with f is

$$(g * f)(x) = \int g(y)f(x-y)\,dy. \tag{8.27}$$

Theorem 8.7 *If g is in L^1 and f is in L^p, the $g * f$ is in L^p. Furthermore,*

$$\|g * f\|_p \leq \|g\|_1 \|f\|_p. \tag{8.28}$$

Proof: Think of the convolution as being the weighted integral of the functions f_y, where $f_y(x) = f(x-y)$. Thus

$$g * f = \int g(y) f_y\,dy \tag{8.29}$$

as an equation in L^p. Thus

$$\|g * f\|_p = \| \int g(y) f_y\,dy\|_p \leq \int |g(y)| \|f_y\|_p\,dy = \int |g(y)| \|f\|_p\,dy. \tag{8.30}$$

Theorem 8.8 *Fix $1 \leq p < \infty$. If δ_ϵ for $\epsilon > 0$ is a family of approximate delta functions obtained by scaling a positive function, then for each f in L^p*

$$\|\delta_\epsilon * f - f\|_p \to 0 \tag{8.31}$$

as $\epsilon \to 0$.

Proof: We have

$$(\delta_\epsilon * f)(x) = \int \delta_\epsilon(y) f(x-y)\,dy = \int \delta_1(z) f(x - \epsilon z)\,dz. \tag{8.32}$$

Hence

$$\delta_\epsilon * f - f = \int \delta_1(z)[f_{\epsilon z} - f]\,dz. \tag{8.33}$$

It follows that

$$\|\delta_\epsilon * f - f\|_p \leq \int \delta_\epsilon(y) \|f_{\epsilon z} - f\|_p\,dy. \tag{8.34}$$

The continuity of translation in L^p implies that for each z the norm $\|f_{\epsilon z} - f\|_p$ approaches zero as $\epsilon \to 0$ The dominating function $\delta_1(z) 2\|f\|_p$ is independent of ϵ. Therefore the last integral goes to zero by the dominated convergence theorem.

Notice that if δ_ϵ is a smooth approximate delta function family, then this theorem shows that a function in L^p with $1 \leq p < \infty$ can be approximated in the L^p sense by smooth functions. On the other hand, this is certainly not true for L^∞. This is because translation is not continuous in L^∞.

Furthermore, here is a technical observation. The convergence of $\delta_\epsilon * f$ to f in L^p is far from being uniform on bounded sets in L^p.

8.3 Sobolev spaces

Let U be an open subset of $\mathbf{R}^n$. The space $W^{k,p}(U)$ consists of all functions u on U such that for all $|\alpha| \leq k$ the derivative $D^\alpha u$ is in $L^p(U)$ (as a weak derivative or distribution derivative). This is the Banach space of functions in $L^p(U)$ whose derivatives up to order k are also in $L^p(U)$. The norm on this space is

$$\|u\|_{W^{k,p}} = \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|_p \right)^{\frac{1}{p}}. \tag{8.35}$$

Thus convergence of functions in the Sobolev space means convergence in L^p for the function and all derivatives up to order k. For $W^{k,\infty}$ the definition of the norm has to be modified in the appropriate way

The space $H^k(U)$ is the space $W^{k,2}(U)$. This is a particularly important special case because it is a Hilbert space. It is the Hilbert space of functions in $L^2(U)$ whose derivatives up to order k are also in $L^2(U)$.

Of these, perhaps the most important of all is the space $H^1(U)$. This is the Hilbert space of all functions u in L^2 such that the first partial derivatives are also in $L^2(U)$. It is a natural space in the context of energy arguments.

Recall the explicit definition of distribution derivative. The statement that $D^\alpha u = v$ in the distribution or weak sense means that

$$\int_U u D^\alpha \phi \, dx = (-1)^{|\alpha|} \int_U v \phi \, dx \tag{8.36}$$

for all functions ϕ in $C_c^\infty(U)$.

In order to see how this definition works, let us verify the fact that the Sobolev space $W^{k,p}(U)$ is a Banach space. The only question that is not quite routine is to show that it is a complete metric space.

Theorem 8.9 *The Sobolev space $W^{k,p}(U)$ is a complete metric space with respect to its norm. Thus it is a Banach space of functions.*

Proof: Suppose that u_m is a Cauchy sequence in $W_{k,p}(U)$. This means that for each α with $|\alpha| \leq k$ we have $D^\alpha u_m$ as a Cauchy sequence in $L^p(U)$. Since $L^p(U)$ is known to be a Banach space, it follows that there is a function v_α such that $D^\alpha u_m \to v_\alpha$ in the $L^p(U)$ sense. From this we see that for every ϕ in $C_c^\infty(U)$ we have

$$\int_U D_\alpha u_m \phi \, dx \to \int_U v_\alpha \phi \, dx. \tag{8.37}$$

In particular, u_m is a Cauchy sequence in $L^p(U)$, and u_m converges to a function u in $L^p(U)$. It follows that

$$\int_U D_\alpha u_m \phi \, dx = (-1)^{|\alpha|} \int_U u_m D^\alpha \phi \, dx \to (-1)^{|\alpha|} \int_U u D^\alpha \phi \, dx. \tag{8.38}$$

This shows that

$$(-1)^{|\alpha|}\int_U uD^\alpha\phi\,dx = \int_U v_\alpha\phi\,dx \tag{8.39}$$

for each test function ϕ. In other words, $D^\alpha u = v_\alpha$. Therefore $D^\alpha u_m \to D^\alpha u$. This is enough to show that u_m converges to u in $W^{k,p}(U)$.

Functions in a Sobolev space need not be smooth; in higher dimensions they need not even be continuous. However it is a remarkable fact that they can be approximated in the Sobolev norm by smooth functions, at least when $1 \le p < \infty$.

The following theorem shows that functions in Sobolev spaces may be approximated by smooth functions away from the boundary. What happens at the boundary is more delicate, and this topic is not treated here. Anyway, here is the local result.

Theorem 8.10 *Let $1 \le p < \infty$. Let u be in $W^{k,p}(U)$. Let V be an open set that is compactly contained in U, that is, $\bar{V}$ is compact and contained in U. Let δ_ϵ be an approximate delta function that is smooth and has compact support. Then*

$$\|\delta_\epsilon * u - u\|_{W^{k,p}(V)} \to 0. \tag{8.40}$$

Proof: Since δ_ϵ has compact support, the function $\delta_\epsilon * u$ is defined on V for ϵ small enough. Since δ_ϵ is smooth, the function $\delta_\epsilon * u$ is even in $C^\infty(V)$. The task is to show that

$$D^\alpha(\delta_\epsilon * u) \to D^\alpha u \tag{8.41}$$

in $L^p(V)$ for all α with $|\alpha| \le k$.

First we compute that for x in V we have

$$D^\alpha(\delta_\epsilon * u)(x) = \int_U D_x^\alpha \delta_\epsilon(x-y)u(y)\,dy = (-1)^{|\alpha|}\int_U D_y^\alpha \delta_\epsilon(x-y)u(y)\,dy. \tag{8.42}$$

However since $\phi(y) = \delta_\epsilon(x-y)$ is in $C_c^\infty(U)$, we also have

$$(-1)^{|\alpha|}\int_U D_y^\alpha \delta_\epsilon(x-y)u(y)\,dy = \int_U \delta_\epsilon(x-y)D_y^\alpha u(y)\,dy, \tag{8.43}$$

by the definition of distribution derivative. This shows that

$$D^\alpha(\delta_\epsilon * u) = \delta_\epsilon * D^\alpha u \tag{8.44}$$

on V. The result follows from properties of convolution for L^p functions.

8.4 Dirichlet boundary conditions

Let U be an open set. We now want to define the Sobolev space $W_0^{k,p}(U)$ of functions u defined on U with derivatives up to order k in $L^p(U)$, and Dirichlet boundary conditions. We take this to be simply the closure of $C_c^\infty(U)$ in $W^{k,p}(U)$.

Similarly, the Sobolev space $H_0^k(U)$ is the space of functions u defined on U with derivatives up to order k in $L^2(U)$, and with Dirichlet boundary conditions. This is a Hilbert space.

When U is a bounded region and $k \geq 1$, then the space $W_0^{k,p}(U)$ is a proper subspace of $W^{k,p}(U)$. When $p < \infty$ it is obtained, roughly speaking, by imposing k boundary conditions at each boundary point. Thus, for example, functions in $W_0^{1,p}(U)$ must vanish at the boundary. On the other hand, functions in $W_0^{2,p}(U)$ must vanish at the boundary, and their derivatives in the direction normal to the boundary must also vanish at the boundary. Of course, all this is imprecise without a more careful specification of what it means to have values at the boundary. The nice thing about the definition is that it bypasses this question.

It is not too hard to show that for $1 \leq p < \infty$ the spaces $W_0^{k,p}(\mathbf{R}^n) = W^{k,p}(\mathbf{R}^n)$ are equal. In particular, when $p = 2$, we have $H_0^k(\mathbf{R}^n) = H^k(\mathbf{R}^n)$. There is no boundary, and the boundary conditions make no difference.

8.5 The Gagliardo-Nirenberg-Sobolev inequality

Let $1 \leq p < n$. We define the Sobolev conjugate to be the number p^* such that

$$\frac{1}{p^*} + \frac{1}{n} = \frac{1}{p}. \tag{8.45}$$

Thus $p < p^* < \infty$.

For the Sobolev conjugate we will prove the result

$$\|u\|_{p^*} \leq C \|D_x u\|_p. \tag{8.46}$$

If there is to be such a result, then the dimensions must coincide on both sides. Again dx has dimension of length to the nth power, and D_x has dimension length to the -1 power. Thus the integral on the left has dimension n and the integral on the right has dimension $n - 1$. This gives the relation

$$\frac{n}{p^*} = \frac{n-1}{p} \tag{8.47}$$

defining the Sobolev conjugate.

It may be worth spelling out the dimensional analysis more carefully. Suppose that the inequality is true for all u, with the same constant C. Then in particular, for each $a > 0$, it is true for $u_a(z) = u(x/a)$. Thus

$$\left(\int |u(x/a)|^{p^*}\,dx\right)^{\frac{1}{p^*}} \leq C \left(\int |D_x u(x/a)|^p\,dx\right)^{\frac{1}{p}}. \tag{8.48}$$

However using $x = ay$ and $dx = a^n\,dy$, this gives

$$a^{\frac{n}{p^*}} \left(\int |u(y)|^{p^*}\,dy\right)^{\frac{1}{p^*}} \leq C a^{\frac{n-1}{p}} \left(\int |D_y u(y)|^p\,dy\right)^{\frac{1}{p}}. \tag{8.49}$$

If the power were not the same on both sides, then $a > 0$ could be chosen to violate the inequality.

Lemma 8.1 *Let $n > 1$ and define p_1^* with $1 < p_1^* < \infty$ by*

$$\frac{1}{p_1^*} = 1 - \frac{1}{n}. \tag{8.50}$$

Then for all C^1 functions u with compact support we have

$$\|u\|_{p_1^*} \leq \frac{1}{2}\|Du\|_1. \tag{8.51}$$

Proof: It is easy to prove from the fundamental theorem of calculus that for each i we have $|u(x)| \leq (1/2)g_i(x)$, where

$$g_i(x) = \int_{-\infty}^{\infty} |D_{x_i}u(x)|\,dx_i. \tag{8.52}$$

This implies that

$$|u(x)| \leq \frac{1}{2}\prod_{i=1}^{n} g_i(x)^{\frac{1}{n}}. \tag{8.53}$$

Now apply Hölder's inequality to each of the one dimensional integrals over dx_i. For each such integral we have $n-1$ factors, and $1/p_1^* = (n-1)/n$. We obtain

$$\|u\|_{p_1^*} \leq \frac{1}{2}\prod_{i-1}^{n} \|D_{x_i}u(x)\|_1^{\frac{1}{n}} \tag{8.54}$$

Since each $\|D_{x_i}u\|_1 \leq \|Du\|_1$, this gives the result.

Theorem 8.11 *Let $n > 1$ and $1 \leq p < n$. Define p^* with $p < p^* < \infty$ by*

$$\frac{1}{p^*} = \frac{1}{p} - \frac{1}{n}. \tag{8.55}$$

Then for all C^1 functions u with compact support we have

$$\|u\|_{p^*} \leq \frac{1}{2}\frac{p^*}{p_1^*}\|Du\|_p. \tag{8.56}$$

Proof: From the lemma applied to $|u|^{\frac{p^*}{p_1^*}}$ we see that

$$\|u\|_{p^*}^{\frac{p^*}{p_1^*}} = \||u|^{\frac{p^*}{p_1^*}}\|_{p_1^*} \leq \frac{1}{2}\|D|u|^{\frac{p^*}{p_1^*}}\|_1 = \frac{1}{2}\frac{p^*}{p_1^*}\||u|^{\frac{p^*}{p_1^*}-1}Du\|_1. \tag{8.57}$$

Apply Hölder's inequality. This gives

$$\|u\|_{p^*}^{\frac{p^*}{p_1^*}} \leq \frac{1}{2}\frac{p^*}{p_1^*}\||u|^{\frac{p^*}{q}}\|_q\|Du\|_p, \tag{8.58}$$

where

$$\frac{1}{q} = 1 - \frac{1}{p} = \frac{1}{p_1^*} - \frac{1}{p^*}. \tag{8.59}$$

Another form for this is

$$\|u\|_{p^*}^{\frac{p^*}{p_1^*}} \le \frac{1}{2}\frac{p^*}{p_1^*}\|u\|_{p^*}^{\frac{p^*}{q}}\,\|Du\|_p. \tag{8.60}$$

This can be rearranged to give the result.

It may help to record that $p^*/p_1^* = p(n-1)/(n-p)$. In particular, when $p = 2$, the coefficient $p(n-1)/(n-p)$ divided by 2 has the value $(n-1)/(n-2) \le 2$ for $n \ge 3$.

8.6 The Poincaré inequality

The theorem obviously extends to u in the space $W_0^{1,p}(U)$, since by definition these functions may be approximated by functions with compact support. This leads to the following Poincaré inequality for the case when U has finite measure.

Corollary 8.3 *Let $n > 1$ and $1 \le p < n$. Define p^* with $p < p^* < \infty$ by*

$$\frac{1}{p^*} = \frac{1}{p} - \frac{1}{n}. \tag{8.61}$$

Let U be an open set with finite measure. Let $1 \le q \le p^*$ and let

$$\frac{1}{q} = \frac{1}{r} + \frac{1}{p^*}. \tag{8.62}$$

Then for all C^1 functions u with compact support in U we have

$$\|u\|_q \le (\mathrm{meas}(U))^{\frac{1}{r}}\frac{1}{2}\frac{p^*}{p_1^*}\|Du\|_p. \tag{8.63}$$

In particular, if $q = p$ and $r = n$, then

$$\|u\|_p \le (\mathrm{meas}(U))^{\frac{1}{n}}\frac{1}{2}\frac{p^*}{p_1^*}\|Du\|_p. \tag{8.64}$$

This corollary implies that when U is an open set of finite measure and $p < n$, then $\|u\|_p$ is bounded by a multiple of $\|Du\|_p$. It follows that the Sobolev norm on $W_0^{1,p}(U)$ is equivalent to the norm $\|Du\|_p$.

In particular, for this case, when U has finite measure, this says that for $2 < n$ the Sobolev norm on $H_0^1(U)$ is equivalent to the norm $\|Du\|_2$.

When $n = 2$ this can be seen in a somewhat different way. If U has finite measure, then $\|Du\|_1$ can be bounded in terms of $\|Du\|_2$. On the other hand, we know that for $n = 2$ we have $\|u\|_2 \le (1/2)\|Du\|_1$. So again the Sobolev norm on $H_0^1(U)$ is equivalent to the norm $\|Du\|_2$.

Chapter 9

Spectral theory and evolution equations: Discrete spectrum

9.1 Separation of variables

Consider the heat equation

$$\frac{\partial u}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2 u}{\partial x^2} \tag{9.1}$$

for $0 \leq x \leq L$ and $0 \leq t$. The boundary conditions are $u(0,t) = u(L,t) = 0$ and $u(x,0) = g(x)$.

A standard procedure for solving such an equation is separation of variables. One looks for a solution that is a product of a function of t with a function of x. Such a solution is $\exp(-\lambda t)\sin(kx)$ where $\lambda = \sigma^2 k^2/2$. If the solution is to satisfy the spatial boundary conditions, then $k_n = n\pi/L$ for some $n = 1, 2, 3, \ldots$. Correspondingly, $\lambda_n = \sigma^2 k_n^2/2$. However this only gives a rather special kind of initial condition.

We can get a general initial condition by expanding

$$g(x) = \sum_n c_n \sin(k_n x) \tag{9.2}$$

in a Fourier sine series. Then the solution is the corresponding expansion

$$u(x,t) = \sum_n c_n \exp(-\lambda_n t)\sin(k_n x). \tag{9.3}$$

This representation of the solution is quite useful. Each n is thought of as indexing a decay mode of the solution. The mechanism of decay, of course, is the smoothing out of the solutions and ultimate absorption at the boundary. The larger n values correspond to larger λ_n and hence to faster decay. Thus the decay is soon dominated by the first few modes, even by the first mode. The rate of decay is also determined by the first mode. For the present problem this is $\lambda_1 = \sigma^2\pi^2/(2L^2)$. The decay is more rapid if the diffusion constant is large or the length of the interval is short.

Consider the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \tag{9.4}$$

for $0 \le x \le L$. The boundary conditions are $u(0,t) = u(L,t) = 0$. The initial conditions are $u(x,0) = g(x)$ and $\partial u/\partial t(x,0) = h(x)$.

Again this can be solved by separation of variables. Possible product solutions are $\cos(\omega_n t)\sin(k_n x)$ and $\sin(\omega_n t)\sin(k_n x)$. Here $\omega_n = ck$. If the solution is to satisfy the spatial boundary conditions, then $k_n = n\pi/L$ for some $n = 1, 2, 3, \ldots$. Correspondingly, $\omega_n = cn\pi/L$.

We can get a general initial condition by expanding both

$$g(x) = \sum_n c_n \sin(k_n x) \tag{9.5}$$

and

$$h(x) = \sum_n d_n \sin(k_n x) \tag{9.6}$$

in a Fourier sine series. Then the solution is the corresponding expansion

$$u(x,t) = \sum_n [c_n \sin(\omega_n t) + d_n \frac{\sin(\omega_n t)}{\omega_n}] \sin(k_n x). \tag{9.7}$$

The interpretation of this solution is best seen by calculating the energy. This is

$$\frac{1}{2}\int_0^L \left[\left(\frac{\partial u}{\partial t}\right)^2 + \left(\frac{\partial u}{\partial x}\right)^2 \right] dx = \frac{1}{2} L \sum_n [\omega_n^2 c_n^2 + d_n^2]. \tag{9.8}$$

Each n is thought of as indexing an oscillation mode of the solution. Each mode represents a standing wave formed by reflection at the boundary. The larger n values correspond to larger ω_n^2 and hence to higher energy. The energy equation shows that each mode conserves energy separately. Higher modes have higher energy.

Both the heat equation and the wave equation solutions depend ultimately on the expansion in sine functions. Why does this work? We may think of the sine functions as eigenvectors of a differential operator. We have

$$-\frac{\partial^2}{\partial x^2} \sin(k_n x) = k_n^2 \sin(k_n x). \tag{9.9}$$

Then we should be able to appeal to a theorem on expansion of an arbitrary vector in terms of eigenvectors.

It is somewhat tricky to make this rigorous for differential operators. It is thus convenient to look at the inverse of the operator, which is an integral operator. If we solve

$$-\frac{\partial^2 u}{\partial x^2} = f \tag{9.10}$$

with boundary condition $u(0) = u(L) = 0$, then we get

$$u(x) = \int_0^x (1 - \frac{x}{L}) y f(y)\, dy + \int_x^L (1 - \frac{y}{L}) x f(y)\, dy. \tag{9.11}$$

This is of the form

$$u(x) = \int_0^L G(x, y) f(y)\, dy, \tag{9.12}$$

where $G(x, y) = (1 - x/L)y$ for $y \leq x$ and $G(x, y) = (1 - y/L)x$ for $y \geq x$. The important thing is that $G(x, y) = G(y, x)$.

This is a symmetric integral operator, the analog of a symmetric matrix. We know that a symmetric matrix has an orthogonal basis of eigenvectors. The same should be true for a symmetric integral operator, at least in special circumstances. In the present case the operator satisfies

$$\int_0^L G(x, y) \sin(k_n y)\, dy = \frac{1}{k_n^2} \sin(k_n x). \tag{9.13}$$

Furthermore, the sine functions are orthogonal in the Hilbert space $L^2([0, L],\, dx)$. Each function has norm squared equal to $L/2$, so there is no problem normalizing them to form an orthonormal basis.

9.2 Spectral theorem for compact self-adjoint operators

Let $\mathcal{H}$ be a Hilbert space. We may as well take the scalars to be real, as this will be sufficient for most of our applications. Let $A : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator. To say the A is bounded is to say that A maps bounded sets into bounded sets. This is equivalent to saying that

$$\|A\| = \sup_{\|u\| \leq 1} \|Au\| < \infty. \tag{9.14}$$

A linear operator is bounded if and only if it is continuous.

A bounded linear operator A is self-adjoint (or symmetric) if $\langle Au, v\rangle = \langle u, Av\rangle$ for all u and v in $\mathcal{H}$. For a self-adjoint operator we can look at the quadratic form $\langle u, Au\rangle$. We have the following result.

Lemma 9.1 *For a bounded self-adjoint operator A we have*

$$\sup_{\|u\| \leq 1} |\langle u, Au\rangle| = \|A\|. \tag{9.15}$$

Proof: The fact that the quadratic form is bounded by the operator norm is obvious from the Schwarz inequality. The other direction takes more work. Let $|\mu|$ be the maximum associated with the quadratic form. We must bound $\|A\|$ by $|\mu|$.

The idea is to use the identity

$$4\langle u, Av\rangle = \langle (u+v), A(u+v)\rangle - \langle (u-v), A(u-v)\rangle \tag{9.16}$$

that is a consequence of symmetry. It follows that

$$4|\langle u, Av\rangle| \leq |\mu|(\|u+v\|^2 + \|u-v\|^2) = 2|\mu|(\|u\|^2 + \|v\|^2) \leq 4|\mu|. \tag{9.17}$$

for arbitrary vectors u and v with length bounded by one. In particular we can take $u = Av/\|Av\|$. This gives $\|Av\| \leq |\mu|$ for arbitrary vector v with length bounded by one. The bound follows.

A standard device for dealing with self-adjoint operators is to look at the quadratic form $\langle u, Au\rangle$ restricted to the unit sphere $\langle u, u\rangle = 1$. Suppose that $|\mu| = \|A\|$. We would like μ to be an eigenvalue of u. Compute the norm

$$\|Au - \mu u\|^2 = \|Au\|^2 - 2\mu\langle u, Au\rangle + \mu^2 \leq 2\mu^2 - 2\mu\langle u, Au\rangle. \tag{9.18}$$

We can take a sequence u_n such that $\langle u_n, Au_n\rangle \to \mu$ as $n \to \infty$. It follows that these form approximate eigenvectors of A, in the sense that $Au_n - \mu u_n$ approaches zero as n tends to infinity. However there is no guarantee that the u_n converge.

The problem is the following. In the finite dimensional situation the unit sphere is compact. A continuous function defined on a compact set assumes maximum and minimum values at points in the set. However in infinite dimensions the unit sphere in Hilbert space is not compact. So the argument fails, and there are many cases when the maximum or minimum is not assumed.

It is thus false in general that a bounded (that is, continuous) linear operator must have an eigenvector in the Hilbert space. However certain special operators do provide the compactness that is needed. That is the subject of the following discussion.

A subset S of a metric space X is totally bounded if for every $\epsilon > 0$ there exists a finite set of ϵ balls that cover S. Clearly a totally bounded set is bounded.

Theorem: A subset of a complete metric space is compact if and only if it is closed and totally bounded.

In a finite dimensional Hilbert space every bounded set is totally bounded. But in an infinite dimensional Hilbert space the unit ball is not totally bounded. This is because the unit basis vectors are all a distance $\sqrt{2}$ from each other. Thus for $\epsilon < \sqrt{2}$ there is no cover of the unit basis vectors by finitely many ϵ balls.

A linear operator is said to be compact if it maps bounded sets into compact sets. Clearly a compact operator is also a bounded operator.

Lemma 9.2 *Let A be a compact self-adjoint operator. Then there is a real number μ with $|\mu| = \|A\|$ that is an eigenvalue of A.*

Proof: We have shown that there is a sequence u_n such that $Au_n - \mu u_n$ tends to zero. By compactness, there is a subsequence (call it again u_n) such that Au_n converges. But then (supposing $\mu \neq 0$) it follows that u_n converges. It follows by continuity that $Au = \mu u$.

Theorem 9.1 *Let A be a compact self-adjoint operator acting in $\mathcal{H}$. Then there is an orthonormal basis for $\mathcal{H}$ consisting of eigenvectors u_k with real eigenvalues μ_k satisfying*

$$Au_k = \mu_k u_k. \tag{9.19}$$

The eigenvalues μ_k approach zero as $k \to \infty$.

The theorem is proved by constructing μ_1 as the eigenvector obtained from the quadratic form $\langle u, Au\rangle$ as in the lemma. Then consider the Hilbert space consisting of all vectors in $\mathcal{H}$ orthogonal to u_1. Again A restricted to this space is a compact self-adjoint operator. So one can repeat the argument in this space and get an eigenvalue μ_2 with $|\mu_2| \leq |\mu_1|$. Continuing in this way one gets eigenvalues μ_k with absolute values that are continually getting smaller. The corresponding eigenvectors u_k form an orthonormal sequence.

We claim that the μ_k constructed in this way satisfy $\mu_k \to 0$ as $k \to \infty$. Otherwise, the sequence $(1/\mu_k)u_k$ would be a bounded sequence of vectors. It would follow from compactness of A that u_k has a convergent sequence. However this is clearly not going to work for an orthonormal sequence.

So far we have constructed an orthonormal sequence of eigenvectors u_k with eigenvalues μ_k approaching zero. All of these non-zero eigenvalues have finite multiplicity. However it is not necessarily true that we have a basis. The reason is that 0 could be an eigenvalue. If we complete this orthonormal set to a basis, then the operator A is zero on all these additional vectors. So this gives the result.

Often 0 will not be an eigenvalue, but in the setting of the theorem it is possible that 0 is an eigenvalue, perhaps even of infinite multiplicity. In most of our applications there will only be non-zero eigenvalues.

Corollary 9.1 *Let A be a compact self-adjoint operator acting in $\mathcal{H}$. Then there is a unitary operator U from ℓ^2 to $\mathcal{H}$ and a diagonal matrix M such that*

$$AU = UM. \tag{9.20}$$

This is just a restatement of the theorem. Let M be the diagonal matrix with entries μ_k on the diagonal. Let

$$Uf = \sum_k f_k u_k. \tag{9.21}$$

Then

$$AUf = \sum_k f_k Au_k = \sum_k f_k \mu_k u_k = \sum_k (Mf)_k u_k = UMf. \tag{9.22}$$

The importance of this corollary is twofold. It reminds us that finding an orthonormal basis of eigenvectors is the same as diagonalizing with unitary operators. Second, it points the way to an important generalization that works even in the case of continuous spectrum.

9.3 Hilbert-Schmidt operators

Let $k(x, y)$ satisfy

$$\int \int |k(x, y)|^2 \, dx \, dy < \infty. \tag{9.23}$$

Define the operator $K : L^2 \to L^2$ by

$$(Kf)(x) = \int k(x, y) f(y) \, dy. \tag{9.24}$$

Then K is said to be a Hilbert-Schmidt integral operator.

Theorem 9.2 *A Hilbert-Schmidt integral operator is a bounded operator from L^2 to L^2. Furthermore*

$$\|K\| \leq \|k\|_2. \tag{9.25}$$

Note that this is usually a strict inequality. The proof is to use the Schwarz inequality to show that

$$|Kf(x)| \leq \sqrt{\int |k(x, y)|^2 \, dy} \|f\|_2. \tag{9.26}$$

Square, integrate, take the square root. This gives

$$\|Kf\| \leq \|k\|_2 \|f\|_2. \tag{9.27}$$

where the L^2 norm of k is given by a double integral.

Theorem 9.3 *A Hilbert-Schmidt integral operator is a compact operator from L^2 to L^2.*

The proof of this theorem depends on the following lemma.

Lemma 9.3 *If each A_n is a compact operator, and if $\|A_n - A\| \to 0$ as $n \to \infty$, then A is a compact operator.*

The lemma is not difficult to prove. Let B be a bounded set. Let $\epsilon > 0$. Pick n so large that $\|A_n u - Au\| \leq \epsilon/3$ for all u in B. Since the set of all $A_n u$ for u in B is totally bounded, there are finitely many u_k in B such that every $A_n u$ for u in B is within $\epsilon/3$ of some $A_n u_k$. Since $Au - Au_k = Au - A_n u + A_n u - A_n u_k + A_n u_k - Au_k$, it follows that every Au for u in B is within ϵ of some Au_k.

The proof of the theorem is then not difficult. Let functions $u_n(x)$ form an orthogonal basis for $L^2(dx)$. Then $u_n(x)u_m(y)$ form an orthogonal basis for $L^2(dx\,dy)$. We can expand

$$k(x,y) = \sum_n \sum_n c_{nm} u_n(x) u_m(y). \tag{9.28}$$

Let

$$k_N(x,y) = \sum_{n \leq N} \sum_{m \leq N} c_{nm} u_n(x) u_m(y). \tag{9.29}$$

Let K_N the Hilbert-Schmidt operator given by k_N. Then K_N is clearly compact, since its range is finite dimensional. However since the series converges in $L^2(dx\,dy)$, it also follows that $\|K_N - K\| \to 0$ as $N \to \infty$.

If also $K(x,y) = K(y,x)$, then the Hilbert-Schmidt operator is an example of a self-adjoint compact operator.

9.4 Compact embedding of Hilbert spaces

Theorem 9.4 *Let f be a function on $\mathbf{R}^n$ that is bounded and in L^2. Then for each $s > 0$ the operator $f(-\triangle + 1)^{-\frac{s}{2}}$ is a compact operator acting in $L^2(\mathbf{R}^n)$.*

Proof: Let g be a function with Fourier transform $\hat{g}(k) = (k^2+1)^{-\frac{s}{2}}$. Let g_m be the function with Fourier transform equal to $\hat{g}(k)$ for $|k| \leq m$ and zero otherwise. Then $\hat{g}_m$ is in L^2 and so g_m is also in L^2. It follows that the integral operator with kernel $f(x)g_m(x-y)$ is a Hilbert-Schmidt integral operator. In particular it is compact.

Since f is bounded and since $\hat{g}_m$ tends to $\hat{g}$ uniformly as m tends to infinity, the integral operator with kernel $f(x)g_m(x-y)$ tends to the integral operator with kernel $f(x)g(x-y)$ in the uniform norm. It follows that this is also a compact operator. Since this kernel defines the operator specified in the theorem, this completes the proof.

Corollary 9.2 *Let U be an open set with finite measure, and let $H_0^s(U)$ be the Sobolev space of all functions in $L^2(U)$ with $s > 0$ derivatives in $L^2(U)$ and vanishing at the boundary of U. Then the injection of $H_0^s(U)$ into $L^2(U)$ is compact. In other words, every set of functions that is bounded in the $H_0^s(U)$ norm belongs to a set that is compact in the $L^2(U)$ norm.*

The proof of the corollary is given by taking a function f that is 1 on U and zero elsewhere. Since U has finite measure, the function f is in $L^2(\mathbf{R}^n)$. If u is in the Sobolev space, we can write

$$u = fu = f(-\triangle + 1)^{-\frac{s}{2}}(-\triangle + 1)^{\frac{s}{2}}u. \tag{9.30}$$

If the functions u are bounded in the Sobolev space, then by definition the functions $(-\triangle + 1)^{\frac{s}{2}}u$ are bounded in L^2. The result then follows from the compactness of the operator described in the theorem above.

9.5 Positive quadratic forms

Consider a Hilbert space $\mathcal{H}^1$ with inner product $\langle u, v\rangle_1$. The Riesz representation theorem says that every continuous linear functional L on $\mathcal{H}^1$ is represented by a vector in the space. Thus there is a vector u in the Hilbert space with $L(v) = \langle u, v\rangle_1$ for all v in the Hilbert space.

Furthermore, the vector u given by the Riesz representation theorem is given by minimizing the function

$$E[u] = \frac{1}{2}\|u\|_1^2 - L(u). \tag{9.31}$$

Let U be an open set of finite measure. The Dirichlet problem

$$-\triangle u = f \tag{9.32}$$

with u vanishing on ∂U can be solved writing the problem in the weak form

$$\int_U Du \cdot Dv \, dx = \int_U fv \, dx. \tag{9.33}$$

Here u and v belong to the Sobolev space $H_0^1(U)$. This is the space of functions in $L^2(U)$ with one derivative in $L^2(U)$ that can be approximated by functions with compact support in the interior of U. The function f is in $L^2(U)$.

Since the measure of U is finite, the Poincaré inequality says that

$$\int_U u^2 \, dx \leq C^2 \int_U |Du|^2 \, dx \tag{9.34}$$

for u in $H_0^1(U)$. This shows that the norm defined by the L^2 norm of the gradient is equivalent to the usual Sobolev norm defined by the L^2 norm of the function plus the L^2 norm of the gradient.

It follows that $L(v) = \int_U fv \, dx$ is a continuous linear functional on $H_0^1(U)$ and hence is represented by a vector u in $H_0^1(U)$. Furthermore, the vector is obtained by minimizing the energy

$$E[u] = \frac{1}{2}\int_U |Du|^2 \, dx - \int_U fu \, dx. \tag{9.35}$$

This proves the following theorem.

Theorem 9.5 *Suppose that U is an open set with finite measure. Then for each f in $L^2(U)$ the Dirichlet problem*

$$-\triangle u = f \tag{9.36}$$

with u vanishing on ∂U has a unique weak solution u in the Sobolev space $H_0^1(U)$.

One possible physical interpretation of this theorem is in terms of heat flow. The variable u is temperature, and the function f represents a source of heat. The equation describes an equilibrium where the temperature varies in space but is independent of time. The mechanism of the equilibrium is that the heat flows from the source to the boundary. The condition that the region has finite volume implies that the boundary is close enough so as to be able to absorb all the heat, thus maintaining a steady state.

Denote the inner product in the Sobolev space with a subscript 1. Then the solution u satisfies

$$\langle u, v\rangle_1 = \langle f, v\rangle \tag{9.37}$$

for all v in the Sobolev space. Write the solution described in this theorem as $u = Gf$. Then

$$\langle Gf, v\rangle_1 = \langle f, v\rangle \tag{9.38}$$

for all v in the Sobolev space. Note that G is one-to-one, since if $Gf = 0$, then f is orthogonal to a dense subset of $L^2(U)$, and so f must also be zero.

Lemma 9.4 *Suppose that U is an open set of finite measure. Then the operator G that gives the solution of the Dirichlet problem for $-\triangle$ is a bounded operator from $L^2(U)$ to $L^2(U)$.*

Proof: We know from the Poincaré inequality that for an open set of finite measure the $L^2(U)$ norm is bounded by a multiple C of the $H^1_0(U)$ norm. Thus

$$\|Gf\|^2 \le C^2\|Gf\|_1^2 = C^2\langle f, Gf\rangle \le C^2\|f\|\|Gf\|. \tag{9.39}$$

We see that $\|Gf\| \le C^2\|f\|$.

Lemma 9.5 *Suppose that U is an open set of finite measure. Then the operator G that gives the solution of the Dirichlet problem for $-\triangle$ is a compact operator from $L^2(U)$ to $L^2(U)$.*

Proof: Compute

$$\|Gf\|_1^2 = \langle f, Gf\rangle \le \|f\|\|Gf\| \le C^2\|f\|^2. \tag{9.40}$$

Therefore if f is bounded in $L^2(U)$, then Gf is bounded in $H^1_0(U)$, and therefore Gf is compact in $L^2(U)$.

Lemma 9.6 *Suppose that U is an open set of finite measure. Then the operator G that gives the solution of the Dirichlet problem for $-\triangle$ is a self-adjoint operator from $L^2(U)$ to $L^2(U)$.*

Proof: Compute

$$\langle f, Gh\rangle = \langle Gf, Gh\rangle_1 = \langle Gf, h\rangle. \tag{9.41}$$

We can conclude that there is an orthonormal basis of $\mathcal{H}$ consisting of eigenvectors of G. Let $\mu_1 \ge \mu_2 \ge \mu_3 \ge \cdots$ be the eigenvalues. They all satisfy $\mu_k > 0$.

Now we can define L to be the operator defined on the range of G by $LGf = f$. Then L is an operator that may be considered as a definition of $-\triangle$ that is guaranteed to produce values in $L^2(U)$. It follows that there is an orthonormal basis of $\mathcal{H}$ consisting of eigenvectors of L. The corresponding eigenvalues are $\lambda_k = 1/\mu_k > 0$.

Theorem 9.6 *Suppose that U is an open set with finite measure. Then the operator*

$$L = -\triangle \tag{9.42}$$

with Dirichlet boundary conditions has the property that there is an orthonormal basis of $\mathcal{H}$ consisting of eigenvectors of L with eigenvalues $\lambda_k > 0$, each of finite multiplicity, and increasing to infinity as k tends to infinity.

9.6 Evolution equations

Consider the heat equation

$$\frac{\partial u}{\partial t} = \frac{\sigma^2}{2}\triangle u \tag{9.43}$$

for x in an open set U and $t \geq 0$. Suppose that U has finite measure and that Dirichlet boundary conditions are imposed on ∂U. Take the initial condition to be $u(x, 0) = g(x)$.

We know that there is a basis of eigenvectors such that

$$-\frac{\sigma^2}{2}\triangle u_k = \lambda_k u_k. \tag{9.44}$$

Expand

$$g = \sum_k c_k u_k. \tag{9.45}$$

Then the solution of the equation is

$$u(t) = \sum_k \exp(-\lambda_k t) c_k u_k. \tag{9.46}$$

Here $0 < \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots$. The rate of convergence to equilibrium is exponentially fast and is governed by the lowest eigenvalue λ_1. This number depends on the size of the region. In a larger region, where the boundary is far away, the rate λ_1 will be closer to zero, and the convergence to equilibrium will be slower.

Similarly, consider the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \triangle u \tag{9.47}$$

for x in an open set U. Suppose that U has finite measure and that Dirichlet boundary conditions are imposed on ∂U. Take the initial conditions to be $u(x, 0) = g(x)$ and $v(x, 0) = h(x)$.

We know that there is a basis of eigenvectors such that

$$-c^2 \triangle u_k = \lambda_k u_k. \tag{9.48}$$

Expand

$$g = \sum_k c_k u_k \tag{9.49}$$

and

$$h = \sum_k d_k u_k. \tag{9.50}$$

Then the solution of the equation is

$$u(t) = \sum_k [\cos(\sqrt{\lambda_k} t) c_k + \frac{\sin(\sqrt{\lambda_k} t)}{\sqrt{\lambda_k}} d_k] u_k. \tag{9.51}$$

The $\sqrt{\lambda_k}$ are the angular frequencies of vibration. The slowest frequency $\sqrt{\lambda_1}$ depends on the size of the region. A larger region will admit lower frequency vibrations.

Conclusion: A good self-adjoint spectral theory immediately provides solutions of the evolution equations.

All of this extends immediately if we replace $-\triangle$ by a more general elliptic operator in divergence form

$$L = -\sum_i \sum_j \frac{\partial}{\partial x_i} a_{ij}(x) \frac{\partial}{\partial x_j} + c(x), \tag{9.52}$$

Here the coefficient matrix $a_{ij}(x)$ is symmetric and satisfies the uniform ellipticity equation, and $c(x) \geq 0$. The quadratic form defined by integration by parts with Dirichlet boundary conditions defines a Sobolev space inner product. Everything goes as before.

Chapter 10

Spectral theory and evolution equations: Continuous spectrum

10.1 Separation of variables

Consider the heat equation

$$\frac{\partial u}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2 u}{\partial x^2} \tag{10.1}$$

for $0 \le x$ and $0 \le t$. The boundary conditions are $u(0,t) = 0$ and $u(x,0) = g(x)$. The new feature is that this is defined on a space interval of infinite length.

Recall that for the case of an interval $0 \le x \le L$ we could get a general initial condition by expanding

$$g(x) = \sum_k c(k)\sin(kx). \tag{10.2}$$

in a Fourier sine series, where $k = n\pi/L$. Since the L^2 norm squared of $\sin(kx)$ is $L/2$, the coefficient is

$$c(k) = \frac{2}{L}\int_0^L \sin(kx)g(x)\,dx. \tag{10.3}$$

The problem is that these formulas do not have good limits as L goes to infinity.

This can be fixed by changing the normalizations. Let instead

$$g(x) = \sum_k c(k)\sin(kx)\frac{2}{L}, \tag{10.4}$$

where

$$c(k) = \int_0^L \sin(kx)g(x)\,dx. \tag{10.5}$$

Then there is a decent limit. Note that if $k = n\pi/L$, then the spacing is $\Delta k = \pi/L$. Thus the sum becomes an integral

$$g(x) = \frac{2}{\pi}\int_0^\infty c(k)\sin(kx)\,dk, \tag{10.6}$$

where

$$c(k) = \int_0^\infty \sin(kx)g(x)\,dx. \tag{10.7}$$

With the help of this representation we can write the solution of the heat equation as an integral

$$u(x,t) = \frac{2}{\pi}\int_0^\infty \exp(-\lambda_k t)c(k)\sin(kx)\,dk. \tag{10.8}$$

Here $\lambda_k = \sigma^2 k^2/2$. Now the decay modes of the solution are indexed by a continuous parameter.

We can try to put this expansion in the context of the Hilbert space $L^2([0,\infty), dx)$. The initial conditions g should be taken in this space. However now we have a peculiarity. The sine functions appear to be eigenfunctions of a differential operator. We have

$$-\frac{\partial^2}{\partial x^2}\sin(kx) = k^2\sin(k_n x). \tag{10.9}$$

However for fixed k the sine function is not in the Hilbert space. This shows that the linear algebra formulation must be done with some care.

Again it is convenient to look at the inverse of the operator. If we take $a > 0$ and f in L^2 and solve

$$-\frac{\partial^2 u}{\partial x^2} + a^2 u = f \tag{10.10}$$

with boundary condition $u(0) = 0$ and also require that the solution u be in L^2, then we get

$$u(x) = \int_0^x \frac{1}{a}\sinh(ay)\exp(-ax)f(y)\,dy + \int_x^\infty \frac{1}{a}\sinh(ax)\exp(-ay)f(y)\,dy. \tag{10.11}$$

This is of the form

$$u(x) = \int_0^L G(x,y)f(y)\,dy, \tag{10.12}$$

where $G(x,y) = \sinh(ay)\exp(-ax)/a$ for $y \le x$ and $G(x,y) = \sinh(ax)\exp(-ay)/a$ for $y \ge x$. In particular $G(x,y) = G(y,x)$.

This is a symmetric integral operator, the analog of a symmetric matrix. It is bounded as an operator on $L^2([0,\infty),\ dx)$, but it is not compact. We expect that the functions $\sin(kx)$ should play a role in understanding this operator, but since they are not in the Hilbert space, this requires further clarification.

We can make this look like standard linear algebra if we define certain unitary operators. Let

$$(Uc)(x) = \frac{2}{\pi}\int_0^\infty c(k)\sin(kx)\,dk. \tag{10.13}$$

take L^2 functions of wave number $k \geq 0$ to L^2 functions of position $x \geq 0$. This involves an integral of the eigenfunctions over the spectral parameter. The analog of the eigenvalue equation is

$$-\frac{\partial^2}{\partial x^2} Uc = U(k^2 c(k)). \tag{10.14}$$

This is perfectly meaningful as a Hilbert space equation if $k^2 c(k)$ is in L^2.

The same operator U gives the analog of the eigenvalue equation for the integral operator G. The result is

$$GUc = U(\frac{1}{k^2 + a^2} c(k)). \tag{10.15}$$

This is meaningful as a Hilbert space equation if $c(k)$ is in L^2.

The inverse operator is

$$(U^{-1}g)(k) = \int_0^\infty \sin(kx) g(x)\, dx. \tag{10.16}$$

We can represent

$$-\frac{\partial^2}{\partial x^2} g = U(k^2 c(k)) \tag{10.17}$$

where $c = U^{-1}g$, provided that g in L^2 is chosen so that the result is in L^2. It is not unreasonable that this same form of representation also gives the solution

$$u(t) = U(\exp(-\sigma^2 k^2 t/2) c(k)) \tag{10.18}$$

for the solution of the heat equation.

10.2 Continuity lemma for bounded self-adjoint operators

Let A be a bounded self-adjoint operator acting in a Hilbert space $\mathcal{H}$. We shall see that it need not have eigenvectors that belong to $\mathcal{H}$. However there is nevertheless a good spectral theory. The key to this spectral theory is a continuity lemma. In order to get an idea of what this lemma says, it is useful to look at the case when $\mathcal{H}$ is a finite dimensional Hilbert space. In this case there is a complete set of eigenvectors u_n that form an orthonormal basis. We have $Au_n = \lambda_n u_n$. If we take an arbitrary element u of $\mathcal{H}$ and expand it as

$$u = \sum_k c_k u_k, \tag{10.19}$$

then we have the equation

$$\langle u, f(A)u\rangle = \sum_k |c_k|^2 f(\lambda_k). \tag{10.20}$$

In particular, we have

$$\|g(A)u\|^2 = \sum_k |c_k|^2 g(\lambda_k)^2. \tag{10.21}$$

As a corollary, we have

$$\|g(A)\| \le \max_k |g(\lambda_k)| \tag{10.22}$$

This shows that if the $g(\lambda_k)$ are uniformly small, then the operator norm $\|g(A)\|$ is small.

There is no estimate like this for operators that are not self-adjoint. For example, if

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \tag{10.23}$$

and $g(x) = x$, then the only element of the spectrum is zero, but $\|A\| = 1$. This is also related to the fact that there is no good way of taking functions of non-self-adjoint operators. Thus, for instance, the square root function is well defined at zero, but the operator A has no square root.

The key to the general case is a continuity property of the mapping that sends the real polynomial p to the bounded self-adjoint operator $p(A)$.

Lemma 10.1 *Let A be a bounded self-adjoint operator acting in a Hilbert space $\mathcal{H}$. Let K be the interval from $-\|A\|$ to $\|A\|$. Let p be an arbitrary real polynomial. Then*

$$\|p(A)\| \le \sup_K |p|. \tag{10.24}$$

This property says that the map that sends p to $p(A)$ is continuous from $C(K)$ to the space of bounded operators on $\mathcal{H}$. This property is far from evident if one starts from first principles. However it may be proved if we use results from the spectral theory of self-adjoint operators acting in a finite dimensional Hilbert space. (As we have seen, it is obvious for such operators from the usual spectral representation.)

Suppose that the polynomial has degree n. Consider an arbitrary vector u. The finite dimensional Hilbert space consists of the span of the vectors $u, Au, A^2u, \ldots, A^n u$. The operator A does not leave this subspace invariant. However let E be the orthogonal projection onto this subspace. Then EAE does leave this subspace invariant. Furthermore, since $\|EAE\| \le \|A\|$, each eigenvalue of EAE is in the interval K. The spectral representation in finite dimensional space gives the estimate

$$\|p(EAE)\| \le \sup_K |p|. \tag{10.25}$$

In particular,

$$\|p(EAE)u\| \le \sup_K |p| \, \|u\|. \tag{10.26}$$

However $p(EAE)u = p(A)u$. Therefore

$$\|p(A)u\| \le \sup_K |p| \, \|u\|. \tag{10.27}$$

Since u is arbitrary, this proves the estimate of the lemma.

Corollary 10.1 *The mapping defined on polynomials extends by continuity to a mapping that sends the real continuous function f in $C(K)$ to the corresponding self-adjoint operator $f(A)$. This mapping preserves the algebraic operations of addition and multiplication.*

10.3 Spectral theorem for bounded self-adjoint operators

The spectral theorem for compact self-adjoint operators may be stated in the following way. Every compact self-adjoint operator is unitarily equivalent to a diagonal operator. (The compactness implies not only the discreteness but also that the diagonal entries cluster only at zero.)

The spectral theorem for bounded self-adjoint operators may be stated in various equivalent ways. Perhaps the most simple statement is the following: Every bounded self-adjoint operator is unitarily equivalent to a multiplication operator. (The boundedness implies that the multiplication operator is multiplication by a bounded function.)

Theorem 10.1 *Let A be a bounded self-adjoint operator acting in $\mathcal{H}$. Then there is a space $L^2(K, \nu)$ and a real measurable function μ on K and a unitary operator $U : L^2(K, \nu) \to \mathcal{H}$ such that*

$$AU = UM. \tag{10.28}$$

Here M is the operator that multiplies functions in $L^2(K, \nu)$ by the function μ.

Example: Let $a \neq 0$. Consider the bounded self-adjoint integral operator G acting in $L^2([0, \infty), dx)$ given by inverting $(-d^2/dx^2 + a^2)$ while imposing Dirichlet boundary conditions. Then G is isomorphic to multiplication by the bounded real function $\mu(k) = 1/(k^2 + a^2)$.

The spectral theorem says that A is equivalent to M. The matrix M is the analog of a diagonal matrix, but rather than acting in a discrete space ℓ^2 it acts in a possibly continuous space L^2. Of course the theorem contains the discrete situation as a special case if we take the measure ν to be a discrete measure.

The advantage of this formulation is that the operator U is constructed from some kind of eigenvectors of A that do not belong to the Hilbert space, but we never have to speak of these eigenvectors. If f is sharply peaked as a kind of approximate delta function, then Uf is approximately an eigenvector, and the equation $AUf = UMf$ is a kind of approximate eigenvalue equation.

It is important to realize that the set K and the measure ν are not unique. It is the values of the function μ that play the role of the eigenvalues. The set K corresponds to a way of labeling the eigenvalues, and so it can be chosen for convenience, just so long as the values $\mu(k)$ for k in K occur with the appropriate multiplicities.

Let us look at this reparameterization in more detail. Suppose we have a spectral representation in $L^2(K,\nu)$, where the operator is multiplication by μ. Let $r = \phi(k)$, where ϕ is a one-to-one measurable correspondence between K and R. Let ν_ϕ be the image of the measure ν under ϕ. The image measure is defined so that

$$\int_R |f(r)|^2 \, d\nu_\phi(r) = \int_K |f(\phi(k))|^2 \, d\nu(k). \tag{10.29}$$

Let μ_ϕ be the function given by $\mu_\phi(r) = \mu(\phi^{-1}(r))$. The representation with measure ν_ϕ and multiplication operator given by μ_ϕ is isomorphic to the original representation with measure ν and multiplication operator given by μ. The isomorphism W from $L^2(R,\nu_\phi)$ to $L^2(K,\nu)$ is given by $(WF)(k) = f(\phi(k))$. It is easy to check that $\mu W f = W \mu_\phi f$.

Another source of lack of uniqueness comes from changing the weight of the measure. Let w be a measurable function whose values are strictly positive real numbers. The measure ν may be replaced by the measure ν weighted by w. Then we have the identity

$$\int_K |f(k)|^2 \, w(k) \, d\nu(k) = \int_K |\sqrt{w(k)} f(k)|^2 \, d\nu(k). \tag{10.30}$$

Therefore if we define $(Vf)(k) = \sqrt{(w(k)} f(k)$, we get an isomorphism from the L^2 space with the new measure to the L^2 space with the original measure. The multiplication operator in this case remains the same.

Since the spectral representation is not unique, the construction must involve some arbitrary choice. One way to do this is to use cyclic vectors. This may not give the most pleasant representation, but it always work to show that a representation exists.

A vector u is a cyclic vector if the set of all vectors $p(A)u$, where p ranges over polynomials, is dense in the Hilbert space.

Lemma 10.2 *Suppose that the Hilbert space contains a cyclic vector for the operator A. Then the spectral representation may be chosen so that the index set K is the interval from $-\|A\|$ to $\|A\|$, ν is a measure with support in K, and μ is the identity map on K.*

The situation in this lemma corresponds to the case when the multiplicity is one. In that case the eigenvalues can act as their own labels.

Proof of lemma: The lemma relies essentially on the fact, proved above, that we can take arbitrary continuous functions of a bounded self-adjoint operator. Let u be a cyclic vector. Consider the linear function from $C(K)$ to real numbers given by

$$L(f) = \langle u, f(A)u \rangle. \tag{10.31}$$

If $f \geq 0$, then there is a real continuous function g with $f = g^2$. Therefore

$$L(f) = \langle u, g(A)^2 u \rangle = \langle g(A)u, g(A)u \rangle \geq 0. \tag{10.32}$$

Therefore L sends positive functions to positive numbers. This is enough to prove that there is a measure ν with

$$L(f) = \int_K f(k)\, d\nu(k) = \langle u, f(A)u\rangle. \tag{10.33}$$

Now that we have the measure the proof is essentially done. We can define $Ug = g(A)u$. (In particular U sends the function 1 to the cyclic vector u.) If we set $\mu(r) = r$, we have $AUg = Ag(A)u = (\mu g)(A)u = U(\mu g)$. This shows that A acts like multiplication by μ. Furthermore,

$$\int_K g(k)^2\, d\nu(k) = \langle g(A), g(A)u\rangle. \tag{10.34}$$

This shows that U preserves the norm. Therefore U extends by continuity to $L^2(K, \nu)$. In order that $L^2(K, \nu)$ be a genuine Hilbert space it is essential, of course, to identify functions that differ on a set of ν measure zero. The fact that U is onto the whole Hilbert space depends on the fact that u is a cyclic vector.

Example: Let $a > 0$. Consider the bounded self-adjoint integral operator G acting in $L^2([0, \infty), dx)$ given by inverting $(-d^2/dx^2 + a^2)$ while imposing Dirichlet boundary conditions. The construction of the spectral representation given by the lemma involves the choice of a cyclic vector. A convenient choice is to take $u(x) = \exp(-ax)$. From the construction we know that G will be isomorphic to multiplication by r on some space L^2 with measure $\rho(r)\, dr$. The function $\exp(-ax)$ will be isomorphic to the function 1.

If we use instead the spectral representation given by the sine transform, then G is isomorphic to multiplication by the bounded real function $1/(k^2+a^2)$. The sine transform of $\exp(-ax)$ is the function $k/(k^2+a^2)$. The correspondence between the two representations must make r correspond to $1/(k^2 + a^2)$. Furthermore, it must send the function 1 to the function $k/(k^2 + a^2)$. This can be done by sending function $f(r)$ to a function $f(1/(k^2 + a^2))k/(k^2 + a^2)$.

If the two representations are to be isomorphic, we must have

$$\int_0^{\frac{1}{a^2}} f(r)^2 \rho(r)\, dr = \int_0^\infty f(\frac{1}{k^2 + a^2})\frac{k^2}{(k^2 + a^2)^2}\,\frac{2\, dk}{\pi}. \tag{10.35}$$

This is true if $\rho(r) = \sqrt{1/r - a^2}/\pi$. So the construction gives this spectral measure. However the spectral measure given by the sine transform is more convenient.

Proof of theorem: In the general case we can write the Hilbert space as a direct sum of closed subspaces each with its own cyclic vector. Therefore the spectral representation can be taken with K as a disjoint union of copies of the interval from $-\|A\|$ to $\|A\|$. The measure ν is the measure on K that comes from the measures on the individual copies. The function μ restricted to each copy is the identity function.

Remark: Once we have the theorem, we see that it is possible to define $f(A)$ where A is an arbitrary measurable function. No continuity is required!

An excellent reference for the material for this section is Edward Nelson, *Topics in Dynamics I: Flows*, Chapter 5. The chapter is self-contained.

10.4 Positive quadratic forms

If U is an arbitrary open set and if $c > 0$, then the Sobolev space $H_0^1(U)$ can be defined by the quadratic form associated with $-\triangle + c$. This leads to the following theorem.

Theorem 10.2 *Let $c > 0$ be a constant. Suppose that U is an open set. Then for each f in $L^2(U)$ the Dirichlet problem*

$$-\triangle u + cu = f \tag{10.36}$$

with u vanishing on ∂U has a unique weak solution u in the Sobolev space $H_0^1(U)$.

Consider the interpretation in terms of heat flow. The variable u is temperature, and the function f represents a source of heat. The term $-cu$ representations a dissipation that is proportional to the temperature. The equation describes an equilibrium where the temperature varies in space but is independent of time. The mechanism of the equilibrium is that the heat flows from the source to the boundary, and while it is flowing it is also dissipating. The dissipation guarantees that there is a mechanism for absorbing the heat produced by the source, independent of the geometry of the region. If there were no dissipation, then the equilibrium would not be automatic. In fact, we have seen that when the region $U = \mathbf{R}^n$, there is equilibrium without dissipation when $n > 2$, but a gradual buildup of temperature when $n \leq 2$.

Denote the inner product in the Sobolev space with a subscript 1. Then the solution u satisfies

$$\langle u, v\rangle_1 = \langle f, v\rangle \tag{10.37}$$

for all v in the Sobolev space. Write the solution described in this theorem as $u = Gf$. Then

$$\langle Gf, v\rangle_1 = \langle f, v\rangle \tag{10.38}$$

for all v in the Sobolev space. Note that G is one-to-one, since if $Gf = 0$, then f is orthogonal to a dense subset of $L^2(U)$, and so f must also be zero.

Lemma 10.3 *Suppose that U is an open set and $c > 0$. Then the operator G that gives the solution of the Dirichlet problem for $-\triangle + c$ is a bounded operator from $L^2(U)$ to $L^2(U)$. Its norm is bounded by $1/c$.*

Proof: Since the Sobolev norm is here defined with the constant c in the zero order term, we have $c\|u\|^2 \leq \|u\|_1^2$. Thus

$$c\|Gf\|^2 \leq \|Gf\|_1^2 = \langle f, Gf\rangle \leq \|f\|\|Gf\|. \tag{10.39}$$

We see that $c\|Gf\| \leq \|f\|$.

Lemma 10.4 *Suppose that U is an open set and $c > 0$. Then the operator G that gives the solution of the Dirichlet problem for $-\triangle + c$ is a self-adjoint operator from $L^2(U)$ to $L^2(U)$.*

Proof: Compute

$$\langle f, Gh\rangle = \langle Gf, Gh\rangle_1 = \langle Gf, h\rangle. \tag{10.40}$$

We can conclude from the spectral theorem that G is isomorphic to multiplication by a bounded real function μ with $0 < \mu \leq 1/c$ acting on some L^2 space.

Now we can define L to be the operator defined on the range of G by $LGf = f$. Then L is an operator that may be considered as a definition of $-\triangle + c$ that is guaranteed to produce values in $L^2(U)$. It follows that L is isomorphic to multiplication by a real function $\lambda = 1/\mu$. Clearly $c \leq \lambda$. However λ will be unbounded above, so it is important that it is only be defined on the range of multiplication by μ.

Theorem 10.3 *Suppose that U is an open set. Let $c > 0$ be a constant. Then the operator*

$$L = -\triangle + c \tag{10.41}$$

with Dirichlet boundary conditions is isomorphic to multiplication by a real function $\lambda \geq c$.

It follows that the operator $-\triangle$ with Dirichlet boundary conditions is isomorphic to multiplication by a positive function. However we have no guarantee that this function is bounded away from zero. Thus there are two new phenomena due to the presence of a region of infinite measure. Instead of having a discrete family of standing waves, we may have a continuous spectral representation that describes a scattering process. Furthermore, since there is great expanse of space, the waves can be arbitrarily spread out with arbitrarily low frequency.

10.5 Evolution equations

Consider the heat equation

$$\frac{\partial u}{\partial t} = \frac{\sigma^2}{2}\triangle u \tag{10.42}$$

for x in an open set U and $t \geq 0$. Suppose that Dirichlet boundary conditions are imposed on ∂U. Take the initial condition to be $u(x, 0) = g(x)$.

We know that there is a spectral representation for the operator $L = -(\sigma^2/2)\triangle$. That is, there is a space $L^2(K, \nu)$ and a function $\lambda \geq 0$ defined on K such that

$$L = -\frac{\sigma^2}{2}\triangle = U\Lambda U^{-1}. \tag{10.43}$$

Here Λ denotes the operator of multiplying by λ. The domain of L consists of all u in L^2 such that $\lambda U u$ is in L^2.

Let g be in $L^2(U)$ and let $\hat{g} = U^{-1}g$ be in $L^2(K)$. Then the solution of the equation is

$$u(t) = U\exp(-\lambda t)\hat{g}. \tag{10.44}$$

This solution is meaningful for all initial conditions g in the Hilbert space, since it involves multiplication by a bounded function for each $t \geq 0$.

We may think of this in an even more abstract sense. The spectral theorem gives us a way of defining an arbitrary measurable function of an arbitrary self-adjoint operator. Take the operator to be $L \geq 0$. For each $t \geq 0$, the function $\exp(-tx)$ is a bounded measurable function. Thus $\exp(-tL)$ is a bounded operator. The solution of the heat equation for $t \geq 0$ is simply $\exp(-tL)g$.

For an operator that is not self-adjoint it is sometimes possible to define analytic functions of the operator by the use of Taylor series expansions. Note that this does not work in the present case, because $\exp(-tx)$ is not analytic near $x = +\infty$.

Consider the interpretation in terms of heat flow. The variable u represents temperature. We have seen that for equilibrium with a source we may need a dissipative term $-cu$. However with this dissipative term there is always an equilibrium. We are considering at present self-adjoint problems, for which there is a chance of a good spectral theory. Interestingly enough, in order to apply spectral theory for the time-dependent problem, it is enough to be able to solve the equilibrium problem with dissipation. Fortunately, this is easy.

Similarly, consider the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \triangle u \tag{10.45}$$

for x in an open set U. Suppose that Dirichlet boundary conditions are imposed on ∂U. Take the initial conditions to be $u(x, 0) = g(x)$ and $v(x, 0) = h(x)$.

There is a spectral representation for the operator $L = -c^2\triangle$. That is, there is a space $L^2(K, \nu)$ and a function $\lambda \geq 0$ defined on K such that

$$L = -c^2 \triangle = U \Lambda U^{-1}. \tag{10.46}$$

Let $\hat{g} = U^{-1}g$ and $\hat{h} = U^{-1}h$. Then the solution of the equation is

$$u(t) = U[\cos(\sqrt{\lambda}t)\hat{g} + \frac{\sin(\sqrt{\lambda}t)}{\sqrt{\lambda}}\hat{h}]. \tag{10.47}$$

This solution involves multiplication by bounded functions, so it is meaningful for arbitrary Hilbert space initial conditions.

Again we can think of this as just taking functions of an operator. If $L \geq 0$ is a self-adjoint operator, then $\cos(\sqrt{L}t)$ and $\sin(\sqrt{L}t)/\sqrt{L}$ are bounded self-adjoint operators. The solution of the wave equation is thus

$$u(t) = \cos(\sqrt{L}t)g + \frac{\sin(\sqrt{L}t)}{\sqrt{L}}h. \tag{10.48}$$

Conclusion: A good self-adjoint spectral theory immediately provides solutions of the evolution equations. This works even for unbounded regions. However then the spectral representation involves integrals rather than sums.

Again all this extends immediately if we replace $-\triangle$ by a more general elliptic operator in divergence form.

10.6 The role of the Fourier transform

The Fourier transform is a special case of the spectral theorem for bounded self-adjoint operators. However, it is such an important case that it is worth pointing out its exceptional properties.

The Hilbert space is $\mathcal{H} = L^2(\mathbf{R}^n, dx)$. The formulas are much simpler if we choose to think of this as a complex Hilbert space. This helps even if the problem involves only real functions.

The Fourier transform gives an isomorphism from $\mathcal{H}$ to the Hilbert space $L^2(\mathbf{R}^n, dk/(2\pi)^n)$. This gives the spectral representation for a special but very important class of operators.

This includes the operators given by convolutions. If g is an integrable function, then the convolution operator g that sends f into $g * f$ is a bounded operator. The Fourier transform shows that this operator is isomorphic to multiplication by the bounded function $\hat{g}$.

If $g(x) = \overline{g(-x)}$, then convolution by g is a self-adjoint operator. This corresponds to the condition that $\hat{g}(k) = \overline{\hat{g}(k)}$, that is, that $\hat{g}(k)$ is real.

The Fourier transform also gives the spectral representation for operators of translation and differentiation. What is common to all these operators? The essential feature is that they commute with translations. That is, if one translates by a in $\mathbf{R}^n$ and then applies one of these operators, this is the same thing as applying the operator and then translating by a.

Conclusion: The significance of the Fourier transform is that it is a spectral representation that works simultaneously for all translation invariant operators.

Chapter 11

Energy and equilibrium

11.1 Least squares solutions

Let $\mathcal{H}$ be a Hilbert space and let $A : \mathcal{H} \to \mathcal{H}$ be a bounded operator. Let w be a vector in $\mathcal{H}$. Say that we want to solve the equation $Au = w$. It would be nice to be able to do this with a variational principle. If A is not symmetric, then we must use expressions that are quadratic in A and A^*.

Theorem 11.1 *Say that A is a bounded operator and there is a constant $\beta > 0$ such that*

$$\beta\|u\| \leq \|A^*u\|. \tag{11.1}$$

Then for every w in $\mathcal{H}$ there is a solution of $Au = w$. It is obtained by finding the g that minimizes

$$E(v) = \frac{1}{2}\|A^*v\|^2 - \langle w, v\rangle \tag{11.2}$$

and setting $u = A^*g$.

Proof: The hypothesis shows that $\|A^*u\|$ is a norm that is equivalent to the ordinary norm $\|u\|$. Therefore we can apply the Riesz representation theorem to the Hilbert space with this norm and with the linear functional $L(v) = \langle w, v\rangle$. The minimization gives the representing vector g with

$$\langle A^*g, A^*v\rangle = \langle w, v\rangle. \tag{11.3}$$

Remark. This technique amounts to solving

$$AA^*g = w. \tag{11.4}$$

and setting $u = A^*g$. The hypothesis of the theorem implies that the self-adjoint operator AA^* is invertible.

The remaining material of this section is a variant approach to the same problem. The same equation is solved; however the variational principle is slightly different.

Again let $\mathcal{H}$ be a Hilbert space and let $A : \mathcal{H} \to \mathcal{H}$ be a bounded operator. If w is in $\mathcal{H}$, a least squares solution is a vector u in $\mathcal{H}$ that minimizes $\|Au - w\|^2$.

Theorem 11.2 *Say that A is a bounded operator and there is a constant $\beta > 0$ such that*

$$\beta\|u\| \leq \|Au\|. \tag{11.5}$$

Then for every w in $\mathcal{H}$ there is a least squares solution satisfying $A^*Au = A^*w$. It is obtained by finding the u that minimizes

$$E(u) = \frac{1}{2}\|Au\|^2 - \langle w, Au\rangle = \frac{1}{2}\|Au - w\|^2 - \frac{1}{2}\|w|^2. \tag{11.6}$$

Proof: Consider the Hilbert space with norm $\|Av\|$. This is equivalent to the original norm. The linear functional $L(v) = \langle w, Av\rangle$ is continuous, so by the Riesz representation theorem there is a vector u with $\langle Au, Av\rangle = \langle w, Av\rangle$.

Remark: The least squares solution satisfies

$$A^*Au = A^*w. \tag{11.7}$$

The hypothesis of the theorem implies that the self-adjoint operator A^*A is invertible. This gives a formula for the least squares solution.

Corollary 11.1 *Suppose that the hypotheses of the theorem hold, and in addition the adjoint A^* is one-to-one. Then for every w in $\mathcal{H}$ the least squares solution u satisfies $Au = w$.*

11.2 Bilinear forms

We have seen that the Riesz representation theorem says that a continuous linear functional can be represented by the inner product. The Lax-Milgram theorem below is a generalization: It says that a continuous linear function can be represented by a bilinear form. The bilinear form need not be symmetric, but its associated quadratic form must be bounded away from zero.

Theorem 11.3 *Let $\mathcal{H}^1$ be a Hilbert space with inner product $\langle u, v\rangle_1$. Let $\beta > 0$. Suppose that B is a bounded bilinear form on $\mathcal{H}^1$ such that*

$$\beta\|u\|_1^2 \leq B(u, v). \tag{11.8}$$

Let L be a continuous linear functional on the Hilbert space $\mathcal{H}^1$. Then there exists an element u in $\mathcal{H}$ with

$$L(v) = B(u, v). \tag{11.9}$$

Proof: By the Riesz representation theorem there exists w with

$$L(v) = \langle w, v\rangle_1. \tag{11.10}$$

Since B is a bounded bilinear form, it may be represented by a bounded linear operator by

$$B(u, v) = \langle Au, v\rangle_1. \tag{11.11}$$

It is easy to see that $\beta\|u\|_1^2 \leq \langle Au, u\rangle_1 \leq \|Au\|_1\|u\|_1$. Furthermore, $\beta\|u\|_1^2 \leq \langle u, A^*u\rangle_1 \leq \|u\|_1\|A^*u\|_1$. These two estimates together with the result of the last section establish that there is a unique solution of

$$Au = w. \tag{11.12}$$

This proves the theorem.

The proof shows that the solution is given by a two stage variational process. First, the vector w given by the Riesz representation theorem is given by minimizing the energy function

$$E[w] = \frac{1}{2}\|w\|_1^2 - L(w). \tag{11.13}$$

Then the solution u is obtained by minimizing

$$F[u] = \|Au - w\|_1^2. \tag{11.14}$$

This theorem is often used in the following way. The bilinear form is a sum

$$B(u, v) = B_1(u, v) + C(u, v), \tag{11.15}$$

where $B_1(u, v)$ is an inner product on $\mathcal{H}^1$ and $C(u, v)$ is a bounded bilinear form on $\mathcal{H}^1$. If C satisfies the estimate

$$-C(u, u) \leq \alpha B_1(u, u) \tag{11.16}$$

with $\alpha < 1$, then the estimate of the theorem is satisfied with $\beta = 1 - \alpha$. From this point of view, we see that the essential feature is that the negative part of the perturbation be relatively small with respect to the symmetric part given by the inner product.

Sometimes we also have the estimate

$$\pm C(u, u) \leq \alpha B_1(u, u) \tag{11.17}$$

with both signs. In this case we can write

$$B(u, v) = B_1(u, v) + B_1(Ru, v), \tag{11.18}$$

The operator relating B to B_1 is $A = I + R$. The operator R will have norm bounded by $\alpha < 1$, so in this case it is possible to solve $Au = (I + R)u = w$ by a convergent power series expansion in powers of R. Then this solution may be inserted to get $B(u, v) = B_1(w, v)$.

11.3 Equilibrium

Up to now we have been considering self-adjoint operators of the form

$$L_0 = -\sum_i \sum_j \frac{\partial}{\partial x_i} a_{ij}(x) \frac{\partial}{\partial x_j} + c(x). \tag{11.19}$$

We always assume that $a_{ij}(x)$ is a symmetric matrix for each x. We also assume the uniform ellipticity condition. This says that there exists $\theta > 0$ such that for each x the eigenvalues of $a_{ij}(x)$ are bounded below by θ. We suppose that the $b_{ij}(x)$ and $c(x)$ are bounded functions. The second order term in L_0 is a diffusion term, and the zero order term is a dissipative term.

Theorem 11.4 *Consider the symmetric bilinear form*

$$B_0(u, v) = \int_U [\sum_i \sum_j a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + c(x)uv] \, dx \tag{11.20}$$

on $H^1_0(U)$. Assume that the coefficients are bounded and satisfy the uniform ellipticity condition. If $\gamma > 0$ is sufficiently large, then $B_0(u, v) + \gamma\langle u, v\rangle$ defines an inner product on $H^1_0(U)$ that is equivalent to the Sobolev inner product.

Proof: Let the eigenvalues of $a_{ij}(x)$ be bounded below by $\theta > 0$ and above by θ'. Let $c(x)$ be bounded below by c and above by c'. Then it is immediate that

$$B_0(u, u) \leq \int_U [\theta' |Du|^2 + c' u^2] \, dx. \tag{11.21}$$

This shows that the norm given by $B_0 + \gamma$ is bounded above by the Sobolev norm. On the other hand, we have

$$B_0(u, u) \geq \int_U [\theta |Du|^2 + cu^2] \, dx. \tag{11.22}$$

If we take γ so large that $c + \gamma > 0$, this shows that the norm given by $B_0 + \gamma$ is also bounded below by the Sobolev norm.

Now we want to look instead at an operator

$$L = L_0 + \sum_i b_i(x) \frac{\partial}{\partial x_i}. \tag{11.23}$$

The resulting form is

$$B(u, v) = B_0(u, v) + C(u, v), \tag{11.24}$$

where

$$C(u, v) = \sum_i \int_U b_i(x) \frac{\partial u}{\partial x_i} v \, dx. \tag{11.25}$$

This form is no longer symmetric. The physical meaning of the new term is that of transport or drift.

Theorem 11.5 *Let L be the strongly elliptic operator with Dirichlet boundary conditions in U. There exists a sufficiently large dissipation parameter $\gamma > 0$ such that for every f in $L^2(U)$ there is a weak solution of $Lu + \gamma u = f$ in the Sobolev space $H^1_0(U)$.*

This theorem says that for the non-selfadjoint problem with a large dissipation parameter we always have equilibrium. This is not surprising, but it is also not too exciting. We shall see, however, in the chapter on semigroups of operators that this is enough to guarantee solutions of related time dependent problems.

Proof: We formulate this as the problem

$$B_\gamma(u, v) = B(u, v) + \gamma\langle u, v\rangle = \langle f, v\rangle. \tag{11.26}$$

in the Sobolev space $H^1_0(U)$. We need to verify the hypotheses of the Lax-Milgram theorem.

It is easy to bound the absolute value of the transport term in the quadratic form by

$$\pm C(u, u) \leq b \int_U |Du||u|\, dx \leq \frac{1}{2} b \int_U [\epsilon|Du|^2 + \frac{1}{\epsilon}|u|^2]\, dx. \tag{11.27}$$

Here $\epsilon > 0$ is arbitrary, and we can take it so that $b\epsilon = \theta$. Then we get the bound

$$C(u, u) \leq \frac{1}{2}[B_0(u, u) + (\frac{b}{\epsilon} - c)\langle u, u\rangle] \leq \frac{1}{2}[B_0(u, u) + \gamma\langle u, u\rangle], \tag{11.28}$$

where γ is sufficiently large. We can now consider this to be the norm squared for the Sobolev space. Thus we see that $(1/2)[B_0(u, u)+\gamma\langle u, u\rangle] \leq B(u, u)+\gamma\langle u, u\rangle$. So $B(u, v) + \gamma\langle u, v\rangle$ satisfies the hypotheses of the Lax-Milgram theorem.

11.4 Boundedness and compactness

This chapter has treated the strongly elliptic divergence form operator

$$L = -\sum_i \sum_j \frac{\partial}{\partial x_i} a_{ij}(x) \frac{\partial}{\partial x_j} + \sum_i b_i(x) \frac{\partial}{\partial x_i} + c(x) \tag{11.29}$$

This operator is associated with a form $B(u, v)$ so that the equation

$$B_\gamma(u, v) = B(u, v) + \gamma\langle u, v\rangle = \langle f, v\rangle \tag{11.30}$$

is a weak form of the equation $Lu + \gamma u = f$. We have seen that for γ sufficiently large this equation has a solution u in the Sobolev space $H^1_0(U)$ for each f in $L^2(U)$.

Theorem 11.6 *Let L be the strongly elliptic operator with Dirichlet boundary conditions in U. For γ sufficiently large the operator $(L - \gamma)^{-1}$ is bounded from $L^2(U)$ to itself.*

Proof: We write the solution as $u = Gf$, so that we have

$$B_\gamma(Gf, v) = \langle f, v\rangle \tag{11.31}$$

for all v in the Sobolev space. The official definition of L is such that $L + \gamma$ is the inverse of G.

As a consequence we have that

$$B_\gamma(Gf, Gf) = \langle f, Gf\rangle. \tag{11.32}$$

By the preceding estimates and the Schwarz inequality we obtain

$$\beta\|Gf\|_1^2 \le B_\gamma(Gf, Gf) \le \|f\|\|Gf\|. \tag{11.33}$$

Since we can bound $\|Gf\|^2$ by a constant times $\|Gf\|_1^2$, this shows that G is a bounded operator from L^2 to itself.

Theorem 11.7 *Suppose U has finite measure. Let L be the strongly elliptic operator with Dirichlet boundary conditions in U. For γ sufficiently large the operator $(L - \gamma)^{-1}$ is compact from $L^2(U)$ to itself.*

Proof: Suppose that U has finite measure. Then the embedding of $H_0^1(U)$ into $L^2(U)$ is compact. It follows from the inequality above that if f belongs to a bounded set in $L^2(U)$, the Gf belong to a bounded set in $H_0^1(U)$. Consequently, the Gf belong to a compact set in $L^2(U)$. This shows that in this circumstance the operator G is compact.

For operators of this class there is no guarantee that G is self-adjoint. If G is a compact operator, there is some spectral theory. For instance, it is known that the spectrum away from zero consists of eigenvalues of finite multiplicity. However this does not give a complete picture of the structure of the operator.

Theorem 11.8 *Let L be the strongly elliptic operator with Dirichlet boundary conditions in U. For γ sufficiently large the quadratic form of the operator $L+\gamma$ is positive. For $\lambda > \gamma$ we have*

$$\|(L - \lambda)^{-1}\| \le 1/(\lambda - \gamma). \tag{11.34}$$

Proof: If u is a weak solution of $(L + \lambda)u = f$, then we have the identity

$$B(u, v) + \lambda\langle u, v\rangle = \langle f, v\rangle \tag{11.35}$$

for all v in the Sobolev space, and in particular

$$B(u, u) + \lambda\|u\|^2 = \langle f, u\rangle. \tag{11.36}$$

However from the estimate we have

$$0 \le B(u, u) + \gamma\|u\|^2. \tag{11.37}$$

It follows that

$$(\lambda - \mu)\|u\|^2 = \langle f, u\rangle \le \|f\|\|u\|. \tag{11.38}$$

This argument show that $\|(L - \lambda)^{-1}f\| \le 1/(\lambda - \gamma)\|f\|$.

Chapter 12

Semigroup theory and evolution equations

12.1 Exponentials

Suppose that we want to solve a parabolic equation of the form

$$\frac{\partial u}{\partial t} = Au \tag{12.1}$$

where A is a linear differential operator. The initial condition is $u(0) = g$. Formally the solution is

$$u(t) = \exp(tA)g. \tag{12.2}$$

So all it takes is to make sense of the exponential. If A is self-adjoint, then this can be accomplished by spectral theory. For general operators it is sometimes possible to define functions by convergent power series. Unfortunately, if A is unbounded, as is the case for differential operators, the series for the exponential will have very delicate convergent properties at best.

Similarly, suppose that we want to solve a hyperbolic equation of the form

$$\frac{\partial^2 u}{\partial t^2} = -Lu. \tag{12.3}$$

There are two initial conditions: $u(0) = g$ and $\partial u/\partial t(0) = h$. The solution is formally

$$u(t) = \cos(\sqrt{L}t)g + \frac{\sin(\sqrt{L}t)}{\sqrt{L}}h. \tag{12.4}$$

So it appears that we also need trigonometric functions and square root functions of unbounded operators. However this may be reduced to the exponential. We can write the equation as a system

$$\frac{d}{dt}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -L & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}. \tag{12.5}$$

If we write the vector with components u, v as w and the vector with components g, h as z, then this is of the form

$$\frac{dw}{dt} = Aw \tag{12.6}$$

with initial condition $w(0) = z$. The solution is thus

$$w(t) = \exp(tA)z. \tag{12.7}$$

Explicitly, this is

$$\exp(\begin{pmatrix} 0 & 1 \\ -L & 0 \end{pmatrix}) \begin{pmatrix} g \\ h \end{pmatrix} = \begin{pmatrix} \cos(\sqrt{L}t) & \sin(\sqrt{L}t)/\sqrt{L} \\ \sqrt{L}\sin(\sqrt{L}t) & \cos(\sqrt{L}t) \end{pmatrix} \begin{pmatrix} g \\ h \end{pmatrix}. \tag{12.8}$$

12.2 Dissipative operators

The goal is to define $\exp(tA)$ for an operator A that is unbounded but negative in some sense. First we make precise what is meant by negative. The precise concept is dissipative.

First recall that if X is a Banach space, then the dual space X^* is the space of all continuous linear functions from X to the scalars. If μ is in X^* and u is in X, then we write the value of μ on u as $\langle \mu, u \rangle$. This notation emphasizes that the value is bilinear in the two arguments.

In general, if u is an element of the Banach space X, then there exists a non-zero element μ of the dual space X^* such that the value $\langle \mu, u \rangle = \|\mu\|\|u\|$. (This is a consequence of the Hahn-Banach theorem.) Let $D(A)$ be a linear subspace of a Banach space X, and let A be a linear transformation from $D(A)$ to X. The condition that A be dissipative is that for every u in $D(A)$ there exists such a μ with

$$\langle \mu, Au \rangle \leq 0. \tag{12.9}$$

This concept is easiest to understand in the case when X is a Hilbert space. If $\mathcal{H}$ is a Hilbert space, then the Riesz representation theorem says that for each μ in the dual space $\mathcal{H}^*$ there is a w in the space $\mathcal{H}$ such that $\langle \mu, u \rangle = \langle w, u \rangle$ for all u in $\mathcal{H}$. In this equation the bracket on the left denotes evaluation, and the bracket on the right denotes the inner product.

If we want $\langle w, u \rangle = \|w\|\|u\|$, then we must take w to be a positive multiple of u. Thus the condition that A be dissipative takes a simple form in the Hilbert space case: For every u in $D(A)$ the quadratic expression

$$\langle u, Au \rangle \leq 0. \tag{12.10}$$

Lemma 12.1 *Let A be dissipative and take $h > 0$. Consider f in X for which there exists u in $D(A)$ with $(I - hA)u = f$ in $D(A)$. Then*

$$\|(I - hA)^{-1} f\| \leq \|f\|. \tag{12.11}$$

Proof:

$$\|\mu\|\|u\| = \langle \mu, u \rangle \leq \langle \mu, (I - hA)u \rangle \leq \|\mu\|\|(I - hA)u\|. \tag{12.12}$$

12.3 The Hille-Yosida theorem

The idea of the Hille-Yosida theorem is the following. Suppose that A is a dissipative operator. The Laplace operator is the standard example. Suppose one wants to solve the equation

$$\frac{du(t)}{dt} = Au(t) \tag{12.13}$$

forward in time. One can approximate this by a backward difference scheme

$$\frac{u(s+h) - u(s)}{h} = Au(s+h) \tag{12.14}$$

with small $h > 0$. Solving, one gets

$$(I - hA)u(s+h) = u(s). \tag{12.15}$$

If one can solve this implicit equation, then this gives a way of solving the equation approximately. If one wants an approximation to the solution $u(t)$ in terms of $u(0) = g$, then it is a matter of taking $h = t/n$ and iterating n times.

For this to work, one needs stability of the scheme. That is, the solution given by $u(s+h) = (I - hA)^{-1}u(s)$ must not blow up. This is where the assumption that A is negative is important. That makes the operator $(I - hA)^{-1}$ have norm bounded by one.

The fundamental hypothesis of the Hille-Yosida theorem below is that this inverse is defined on the entire Banach space and satisfies

$$\|(I - hA)^{-1}\| \leq 1 \tag{12.16}$$

for $h > 0$. This is sometimes written in the equivalent form $\|(\lambda - A)^{-1}\| \leq 1/\lambda$ for $\lambda > 0$.

For theoretical purposes it is sometimes useful to look at a more complicated difference scheme. Write

$$u(s+h) - u(s) = \frac{1}{\lambda}Au(s+h) + (\frac{1}{\lambda} - h)Au(s). \tag{12.17}$$

The idea is to let h approach zero and then let λ approach infinity. If we solve for $u(s+h)$ in terms of $u(s)$ we obtain

$$u(s+h) = u(s) + hA(I - \frac{1}{\lambda}A)^{-1}u(s). \tag{12.18}$$

Thus the double limit is equivalent to first solving

$$\frac{du(t)}{dt} = A(I - \frac{1}{\lambda}A)^{-1}u(t) \tag{12.19}$$

and then letting $\lambda \to \infty$.

Let

$$A_\lambda = A(I - \frac{1}{\lambda}A)^{-1} = \lambda[(I - \frac{1}{\lambda}A)^{-1} - I]. \tag{12.20}$$

The fact that these two expressions are equivalent is an easy exercise. The advantage of the operator A_λ is that it is a bounded operator that is a good approximation to the unbounded operator A when λ is large. Since it is a bounded operator, there is no difficulty in defining $\exp(tA_\lambda)$ by a convergent power series.

Theorem 12.1 *Let X be a real Banach space. Let $D(A)$ be a dense linear subspace of X, and let $A : D(A) \to X$ be a linear operator. Suppose that for all $\lambda > 0$ the operator $(I - (1/\lambda)A)^{-1}$ is a bounded operator defined on all of X and satisfying*

$$\|(I - \frac{1}{\lambda}A)^{-1}\| \leq 1. \tag{12.21}$$

Then for each u in X and each $t \geq 0$ the limit

$$\exp(-tA)u = \lim_{\lambda \to \infty} \exp(tA_\lambda)u \tag{12.22}$$

exists. Furthermore, for each $t \geq 0$ the norm $\|exp(tA)\| \leq 1$, so $\exp(tA)$ is a contraction. The semigroup identity

$$\exp((t + s)A) = \exp(tA)\exp(sA) \tag{12.23}$$

is satisfied for all $s \geq 0$ and $t \geq 0$. The semigroup is continuous in the sense that for each u in X the function that sends $t \geq 0$ to $\exp(tA)u$ is continuous.

The theorem may be interpreted as giving a definition of the operator $\exp(tA)$ for $t \geq 0$.

Warning: The function that sends $t \geq 0$ to $\exp(tA)$ is not continuous in the operator norm sense. The only exception is when A happens to be a bounded operator.

It is illuminating to write out the power series for $\exp(tA_\lambda)$ explicitly. The result is

$$\exp(tA_\lambda) = \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \, (I - \frac{1}{\lambda}A)^{-n}. \tag{12.24}$$

This gives a concrete description of this solution procedure. Take fixed step size $1/\lambda$. The number of steps n in the iteration is a Poisson random variable with mean λt. The approximate solution is given by taking the expected value of the iterated solution with variable number of steps. Note also that it follows from estimating this series expansion that $\|\exp(tA_\lambda)\| \leq 1$ for each $t \geq 0$.

Lemma 12.2 *For each w in X we have $(I - (1/\lambda)A)^{-1}w \to w$ as $\lambda \to \infty$.*

Proof: From the identity above we see that for w in $D(A)$ we have

$$(I - \frac{1}{\lambda}A)^{-1}w - w = \frac{1}{\lambda}(I - \frac{1}{\lambda}A)^{-1}Aw. \tag{12.25}$$

This gives the result for each w in $D(A)$. However $D(A)$ is dense in X. Since the operators are all bounded by one, the result extends to all w in X.

Lemma 12.3 *For each u in $D(A)$ we have $A_\lambda u \to Au$ as $\lambda \to \infty$.*

Proof: Let u be in $D(A)$. Then

$$A_\lambda u = (I - \frac{1}{\lambda}A)^{-1}Au \to Au \tag{12.26}$$

by the previous lemma.

Proof: In order to prove the convergence of $\exp(tA_\lambda)u$ as $\lambda \to \infty$, we want to compare $\exp(tA_\lambda)$ and $\exp(tA_\mu)$. We have

$$\frac{d}{ds}[\exp((t-s)A_\mu)\exp(sA_\lambda)]u = \exp((t-s)A_\mu)(A_\lambda - A_\mu)\exp(sA_\lambda)u. \tag{12.27}$$

Integrate from 0 to t. This gives

$$\exp(tA_\lambda)u - \exp(tA_\mu)u = \int_0^t \exp((t-s)A_\mu)(A_\lambda - A_\mu)\exp(sA_\lambda)u\,ds. \tag{12.28}$$

It follows from the lemma that for u in $D(A)$ we have

$$\exp(tA_\lambda)u - \exp(tA_\mu)u = \int_0^t \exp((t-s)A_\mu)\exp(sA_\lambda)(A_\lambda u - A_\mu u)\,ds \to 0 \tag{12.29}$$

as λ and μ tend to ∞. Thus by the Cauchy criterion $\exp(tA_\lambda)u$ converges to a limit $\exp(tA)u$ as $\lambda \to \infty$.

Since $\|\exp(tA_\lambda)u\| \leq \|u\|$ for each λ, it follows that $\|\exp(tAu\| \leq \|u\|$. From this it is possible to show that the limit $\exp(tA)u$ of $\exp(tA_\lambda)u$ exists for all u in X. It also shows that the operator norm $\|\exp(tA)\| \leq 1$.

It is then not difficult to verify the semigroup property and the fact that for each u the function that sends $t \geq 0$ to $\exp(tA)u$ is continuous.

Theorem 12.2 *Let X be a real Banach space. Let $D(A)$ be a dense linear subspace of X, and let $A : D(A) \to X$ be a linear operator. Suppose that for all $h > 0$ the operator $(I - hA)^{-1}$ is a bounded operator defined on all of X and satisfying*

$$\|(I - hA)^{-1}\| \leq 1. \tag{12.30}$$

Then for each u in X and each $t \geq 0$ the limit

$$\exp(-tA)u = \lim_{n\to\infty}(I - \frac{t}{n}A)^{-n}u \tag{12.31}$$

exists.

Proof: This is a consequence of the following result.

Theorem 12.3 *Let A be as in the hypotheses of the theorem. Then for each u in X*

$$\| \exp(tA_{\frac{n}{t}})u - (I - \frac{t}{n}A)^{-n}u\| \to 0 \tag{12.32}$$

as $n \to \infty$.

This theorem shows the equivalence of the two definitions. It follows easily from the following quantitative estimate.

Lemma 12.4 *Let A be as in the hypotheses of the theorem. Then for each u in $D(A)$*

$$\| \exp(tA_{\frac{n}{t}})u - (I - \frac{t}{n}A)^{-n}u\| \leq \sqrt{n}\frac{t}{n}\|A_{\frac{n}{t}}u\| \to 0. \tag{12.33}$$

as $n \to \infty$.

This estimate is in turn a consequence of a general fact about contraction operators. The operator $T = (I - (t/n)A)^{-1}$ is a contraction operator.

Lemma 12.5 *Let T be a bounded operator with $\|T\| \leq 1$. Then for each u in X*

$$\| \exp(n(T - I))u - T^n u\| \leq \sqrt{n}\|Tu - u\|. \tag{12.34}$$

as $n \to \infty$.

Proof: We write

$$\exp(n(T - I)) - T^n = \sum_{m=0}^{\infty} \frac{n^m}{m!}e^{-m}\,(T^m - T^n). \tag{12.35}$$

Furthermore, it is easy to see that $\|T^m u - T^n u\| \leq |n - m|\|Tu - u\|$. So

$$\| \exp(n(T - I))u - T^n u\| \leq \sum_{m=0}^{\infty} \frac{n^m}{m!}e^{-m}|m - n|\ \|Tu - u\|. \tag{12.36}$$

This is a mean with respect to the Poisson distribution $(n^m/m!)e^{-m}$, $m = 0, 1, 2, 3, \ldots$. Bound the mean by the root mean square. Thus

$$\sum_{m=0}^{\infty} \frac{n^m}{m!}e^{-m}|m - n| \leq \sqrt{\sum_{m=0}^{\infty} \frac{n^m}{m!}e^{-m}(m - n)^2} = \sqrt{n}. \tag{12.37}$$

The last equality is the standard computation of the standard deviation of a Poisson random variable with mean n. The standard deviation is $\sqrt{n}$. Notice that as the mean gets large, the standard deviation is an increasingly small proportion of the mean. This is what makes the estimate work.

A good reference for this kind of semigroup theory is E. Nelson, *Topics in Dynamics I: Flows*, Chapter 8. This contains a very general theorem of Chernoff that abstracts the idea behind various difference schemes.

Chapter 13

Compactness

13.1 Total boundedness

A metric space E is totally bounded if for every $\epsilon > 0$ there is a finite subset F of E such that every point of E is within ϵ of some point in F. If E is totally bounded, then in particular it is bounded.

Let E be a subset of a metric space X. If X is totally bounded, then so is E. On the other hand, if E is totally bounded, then so is the closure of E in X.

If E is totally bounded, then for every $\epsilon > 0$ there is a finite subset F of E with N points such that the ϵ balls about the points of F cover E. One can ask how N depends on ϵ. The following examples illustrate this point.

Example: Let E be a cube of side L in an n-dimensional Euclidean space. Then the number N of points needed to cover L to within ϵ is bounded by $(CL/\epsilon)^n$ for some constant C. If instead E is a ball of radius L there is a similar estimate. Thus the condition of being bounded and finite dimensional is enough to guarantee total boundedness. However notice that the number of points needed increases rather rapidly with the dimension.

Example. Let E be a ball of radius $L > 0$ in an infinite dimensional Hilbert space, centered at the origin. Take the vectors in an orthonormal basis and multiply them by L. Take $\epsilon = L/2$. It is an easy exercise to show that there is no ϵ approximation of the scaled basis vectors by a finite set. Conclusion: E is not totally bounded.

The quantitative nature of total boundedness may be emphasized by defining the quantity $N(\epsilon)$ to be the number of ϵ balls needed to cover E. Then the ϵ entropy of E is defined to be

$$H(\epsilon) = \log_2 N(\epsilon). \tag{13.1}$$

Here is a result using this notion. Let X be a Banach space and let K be a compact linear operator from X to itself. Let E be the image of the unit ball under K. Then the number of eigenvalues λ of K (counting multiplicity) with $|\lambda| \geq 4\epsilon$ is bounded by $(1/2)H(\epsilon)$. This result of Carl and Triebel is presented in the book Spectral Theory and Differential Operators, by D. E. Edmunds and W. D. Evans.

13.2 Compactness

A metric space E is complete if every Cauchy sequence in E converges to an element of E.

If E is a subset of a complete metric space X, then E is complete if and only if E is closed.

A metric space E is compact if every sequence in E has a subsequence that converges to a point in E. If E is compact, then E is complete.

Suppose E is a subset of a metric space X. If E is compact, then E is closed. On the other hand, if X is compact and E is closed, then E is compact.

Theorem 13.1 *A metric space E is compact if and only if it is complete and totally bounded.*

Proof: First we prove that if E is compact, then it is totally bounded. Suppose that E is not totally bounded. There is some $\epsilon > 0$ such that it is impossible to cover E by epsilon balls. Let f_0 be an arbitrary point in E. Construct inductively a sequence E_n of points of E by the following procedure. Let $B_1, \ldots, B_n$ be ϵ balls about $f_1, \ldots, f_n$. Take f_n outside of the union of these balls. The sequence constructed in this way cannot possibly have a convergent subsequence. This is enough to show that E is not compact.

Suppose, on the other hand, that E is complete and totally bounded. We want to prove that E is compact. Consider a sequence f_n for n in the set $\mathbf{N} = \{0, 1, 2, 3, \ldots\}$ of natural numbers. Suppose each f_n is an element of E. We need to show that this sequence has a subsequence that converges to a point f in E.

Let ϵ_j for $j \geq 1$ be a sequence of strictly positive numbers that decrease to zero. The idea is to construct inductively a decreasing sequence $\mathbf{N}_j$ of subsets of $\mathbf{N}$ and a sequence of balls B_j. Each $\mathbf{N}_j$ is supposed to be infinite. Each ball B_j is to have radius ϵ_j for each $j \geq 1$. Furthermore, the f_n for n in $\mathbf{N}_j$ are supposed to be in B_j.

We start with $\mathbf{N}_0$ equal to $\mathbf{N}$. If $\mathbf{N}_j$ and B_j have been constructed, then the construction of $\mathbf{N}_{j+1}$ and B_{j+1} is as follows. By total boundedness we can cover the space by finitely many balls of radius ϵ_{j+1}. Since $\mathbf{N}_j$ is infinite, there must be an infinite subset $\mathbf{N}_{j+1}$ of $\mathbf{N}_j$ and one of these balls B_{j+1} such that the f_n for n in $\mathbf{N}_{j+1}$ belong to B_{j+1}.

We now can construct the desired subsequence. If $n_0, \ldots, n_j$ have been constructed with f_{n_j} in B_j, then we can find n_{j+1} in $\mathbf{N}_j$ with $n_{j+1} > n_j$. Then $f_{n_{j+1}}$ is in B_{j+1}.

If $k > j$, then f_{n_k} and f_{n_j} are both in the same ϵ_j ball B_j. Thus they are within distance $2\epsilon_j$ from each other. This is enough to show that this subsequence is a Cauchy sequence. Hence by completeness it converges to some f in E. This completes the proof that E is compact.

Corollary 13.1 *A subset of a complete metric space has compact closure if and only if it is totally bounded.*

13.3 The Ascoli-Arzelà total boundedness theorem

Let $C(K)$ be the set of bounded continuous functions on a compact set K (say in Euclidean space). The norm is defined to be the supremum norm; thus the distance between functions is the uniform distance. The Ascoli-Arzelà theorem gives a condition that ensures that a subset E of $C(K)$ be totally bounded.

Example: Let K be the unit interval, and let f_n be the piecewise linear function that is equal to 1 at 0, 0 at $1/n$, and 0 at 1. Let E be the collection of all such functions f_n for $n = 1, 2, 3, \ldots$. Then E is bounded but not totally bounded.

To see that this is true, take $\epsilon = 1/4$. Consider a finite set $\phi_1, \ldots, \phi_k$ of continuous functions. If $\phi_i(0) \leq 3/4$, then the distance of ϕ_i to each f_n exceeds 1/4. So consider the remaining ϕ_j that have $\phi_j(0) > 3/4$. There is an interval $[0, \delta]$ such that on this interval each $\phi_j(x) > 1/2$. However then for $n \geq 1/\delta$ the distance of f_n from each ϕ_k will be greater than 1/4. So for these n the f_n are not approximated by the functions in the finite set.

The problem with this example is that the functions are getting steeper and steeper; in fact they are approximating a function that is not continuous. So we need conditions that control how steep the functions can get. Basically, we need that there is a uniform bound on the derivative. This can be stated more generally as the condition of equicontinuity.

Definition. A family E of functions is bounded at y if the set of numbers $|f(y)|$ for f in E is bounded.

Definition. A family E of functions is equicontinuous at y if for every $\epsilon > 0$ there is a $\delta > 0$ such that for all f in E and all x, $|x - y| < \delta$ implies $|f(x) - f(y)| < \epsilon$.

Theorem 13.2 *Let K be compact and let E be a family of functions in $C(K)$ such that E is bounded at each point and E is equicontinuous at each point. Then E is totally bounded.*

Proof: Let $\epsilon > 0$. The idea of the proof is to find a finite subset A of K and a finite set B of numbers. Then there are finitely many functions ϕ in the set B^A of all functions from A to B. We want to find a finite set F functions in E parameterized by a subset of B^A and such that every function in E is within ϵ of some function in F.

Since E is equicontinuous at each point, for each y in K there is a δ_y such that for all f in E, if x is in the δ_y ball about y, then $f(x)$ is in the $\epsilon/4$ ball about $f(y)$. Since K is compact, there is a finite subset A of K so that the corresponding δ_y balls for y in A cover K.

Since E is bounded at each point, for each y in K the corresponding set of values $f(y)$ for f in E is bounded. Therefore, the set of values $f(y)$ for y in A and f in E is bounded. It follows that there exists a finite set B of numbers such that every such $f(y)$ is within $\epsilon/4$ of some point in E.

Let ϕ be a function from the finite set A to the finite set B. Suppose that there is a function f_ϕ in E such that $f_\phi(y)$ is within $\epsilon/4$ of $\phi(y)$ for each y in A. The finite set F consists of the f_ϕ for ϕ in the finite set B^A.

The remaining task is to show that every function f in E is within ϵ of some f_ϕ in F. Consider f in E. If y is in A, then there is an element $\phi(y)$ of B such that $f(y)$ is within $\epsilon/4$ of $\phi(y)$. This defines the function ϕ.

Consider x in K. There is a y in A such that x is within δ_y of y. Thus $f(x)$ is within $\epsilon/4$ of $f(y)$. We already know that $f(y)$ is within $\epsilon/4$ of $\phi(y)$, and $\phi(y)$ is within $\epsilon/4$ of $f_\phi(y)$. However also $f_\phi(x)$ is within $\epsilon/4$ of $f_\phi(y)$. This reasoning shows that $f(x)$ is within ϵ of $f_\phi(x)$.

Example: Let us take the example where the functions are defined on a cube of length L in n dimensional Euclidean space. The functions have real values in an interval of length M. Furthermore, their derivatives are bounded by a constant C. Consider $\epsilon > 0$. The corresponding δ is ϵ/C. Thus the number of points in A is $(LC/\epsilon)^n$. The number of points in B is (M/ϵ). Thus the number of functions in F is (M/ϵ) to the $(LC/\epsilon)^n$ power. This grows very rapidly as ϵ approaches zero, but it remains finite.

13.4 The Rellich-Kondrachov embedding theorem

We now want to find criteria for compactness in $L^q(U)$, where U is an open subset of Euclidean space.

Lemma 13.1 *Let g be a function such that g and Dg are in L^∞. Let K be a bounded set. Then convolution by g maps bounded sets in L^1 into totally bounded sets in $C(K)$.*

Proof: Suppose that f has bounded L^1 norm. For each fixed f the function

$$u(x) = \int g(x-y)f(y)\,dy \tag{13.2}$$

is continuous. Furthermore, as f varies in the bounded set in L^1, the corresponding u varies in a bounded set in $C(K)$.

We also have

$$Du(x) = \int Dg(x-y)f(y)\,dy. \tag{13.3}$$

This shows that the derivatives Du are uniformly bounded as f varies in L^1. This is enough to prove equicontinuity. The result follows from the Ascoli-Arzelá theorem.

The lemma shows that convolution by the smooth g is a compact operator from $L^1(K)$ to $C(K)$. However on a bounded set L^p convergence implies L^1 convergence. Also, on a bounded set uniform convergence implies L^q convergence. It follows in particular that convolution by g is compact from $L^p(K)$ to $L^q(K)$.

Lemma 13.2 *Let $\delta_1(y) \geq 0$ be zero for $|y| > 1$ and have integral one. Let $\delta_\epsilon(x) = \delta_1(x/\epsilon)/\epsilon^n \geq 0$ be the corresponding approximate delta function family for $\epsilon > 0$. Consider convolution by δ_ϵ as an operator from the Sobolev space $W^{1,1}(\mathbf{R}^n)$ to $L^1(\mathbf{R}^n)$. Then*

$$\|\delta_\epsilon * -I\| \to 0 \tag{13.4}$$

as $\epsilon \to 0$.

Proof: We have

$$(\delta_\epsilon * u)(x) - u(x) = \int \delta_\epsilon(z)[u(x-z) - u(x)]\, dz = \int \delta_1(y)[u(x-\epsilon y) - u(x)]\, dy. \tag{13.5}$$

Hence

$$\int |(\delta_\epsilon * u)(x) - u(x)|\, dx \leq \int \delta_1(y) \int |u(x-\epsilon y) - u(x)|\, dx\, dy. \tag{13.6}$$

Consider the difference quotient

$$u(x-\epsilon y) - u(x) = \int_0^1 \frac{d}{dt} u(x - t\epsilon y)\, dt = -\epsilon \int_0^1 Du(x - t\epsilon y) \cdot y\, dt. \tag{13.7}$$

It follows that as long as $|y| \leq 1$ we have

$$\int |u(x-\epsilon y) - u(x)|\, dx \leq \epsilon \int |Du(x)|\, dx, \tag{13.8}$$

which is independent of y. It follows that

$$\int |\delta_\epsilon * u(x) - u(x)|\, dx \leq \epsilon \int |Du(x)|\, dx. \tag{13.9}$$

So the norm is bounded by ϵ.

Corollary 13.2 *Let $\delta_1(y) \geq 0$ be zero for $|y| > 1$ and have integral one. Let $\delta_\epsilon(x) = \delta_1(x/\epsilon)/\epsilon^n \geq 0$ be the corresponding approximate delta function family for $\epsilon > 0$. Let U be an open set of finite measure. Let $1 \leq p < n$ and let p^* be its Sobolev conjugate with $1/p^* = 1/p - 1/n$. Suppose $1 \leq q < p^*$. Consider convolution by δ_ϵ as an operator from the Sobolev space $W_0^{1,p}(U)$ to $L^q(U)$. Then*

$$\|\delta_\epsilon * -I\| \to 0 \tag{13.10}$$

as $\epsilon \to 0$.

Proof: By the lemma and the fact that U has finite measure we can bound $\|\delta_\epsilon * u - u\|_1$ by a multiple of $\epsilon\|Du\|_p$. On the other hand, we can bound $\|\delta_\epsilon * u - u\|_{p^*}$ by $2\|u\|_{p^*}$ and this in turn by a constant times $\|Du\|_p$ by the Sobolev inequality.

We need to bound $\|\delta_\epsilon * u - u\|_q$. Write $1/q = \theta + (1-\theta)1/p^*$. Applying Hölder's inequality to the product of the θ power and the $1-\theta$ power gives the bound

$$\|\delta_\epsilon * u - u\|_q \leq \|\delta_\epsilon * u - u\|_1^\theta \|\delta_\epsilon * u - u\|_{p^*}^{1-\theta} \leq C\epsilon^\theta \|Du\|_p. \qquad (13.11)$$

So the norm is bounded by a multiple of ϵ^θ. If $q < p^*$ then $\theta > 0$, so this goes to zero with ϵ.

The following is the Rellich-Kondrachov compactness theorem. It follows from the preceding lemmas and corollary.

Theorem 13.3 *Let U be an bounded open set. Let $1 \leq p < n$ and let p^* be its Sobolev conjugate with $1/p^* = 1/p - 1/n$. Suppose $1 \leq q < p^*$. Then the injection from $W_0^{1,p}(U)$ to $L^q(U)$ sends bounded sets into totally bounded sets.*

Proof: Let δ_ϵ be a smooth approximate delta function family with compact supports. Then for each $\epsilon > 0$ convolution by δ_ϵ sends bounded sets in the Sobolev space into totally bounded sets in the L^q space. On the other hand, these operators converge uniformly to the injection. So the injection itself maps bounded sets in the Sobolev space to totally bounded sets in the L^q space.

Corollary 13.3 *Let U be an open set of finite measure. Let $1 \leq p < n$. Then the injection from $W_0^{1,p}(U)$ to $L^p(U)$ sends bounded sets into totally bounded sets.*

13.5 Almost uniform convergence

Say that ϕ is a uniformly continuous function and $f_n \to f$ in L^∞ (that is, uniformly). Then $\phi(f_n) \to \phi(f)$ in L^∞. So this kind of convergence behaves well with respect to non-linear operators.

Say that ϕ is a uniformly continuous function and $f_n \to f$ in the L^P sense for some p with $1 \leq p < \infty$. Then it is not at all clear that $\phi(f_n) \to \phi(f)$ in L^p, except in very special cases. Thus it is of interest to examine the relation of L^p convergence to various kinds of pointwise convergence. In general the pointwise convergence will take place outside of a set of small measure.

Here are two kinds of convergence.

1. Convergence in measure. The sequence f_n converges to f in measure if for every $\epsilon > 0$ the limit as $n \to \infty$ of the measure of the set where $|f_n - f| \geq \epsilon$ is zero.

2. Almost uniform convergence. The sequence f_n converges to f almost uniformly if for every $\delta > 0$ there is an E with $\mu[E] < \delta$ such that f_n converges to f uniformly on the complement of E.

It is evident that almost uniform convergence implies almost everywhere convergence. The converse is true on a finite measure space.

These kinds of convergence behave well under nonlinear operations.

1. If ϕ is uniformly continuous, then $f_n \to f$ in measure implies $\phi(f_n) \to \phi(f)$ in measure.

2. If ϕ is uniformly continuous, then $f_n \to f$ almost uniformly implies $\phi(f_n) \to \phi(f)$ almost uniformly.

Next we examine the relation between various modes of convergence.

Theorem 13.4 *L^p norm convergence for $1 \leq p < \infty$ implies convergence in measure.*

Proof: This is obvious from the Chebyshev inequality

$$\mu[\{x \mid |f_n(x) - f(x)| \geq \epsilon\}] \leq \frac{\|f_n - f\|_p^p}{\epsilon^p}. \tag{13.12}$$

Theorem 13.5 *Convergence in measure implies that a subsequence converges almost uniformly.*

Proof: Suppose that f_n converges to f in measure. Then there is a subsequence f_{n_k} such that the measure of the set where $|f_{n_k} - f| \geq 1/k$ is less than $1/2^{k+1}$. Then the measure of the set $E(m)$ where $\exists k \geq m \, |f_{n_k} - f| \geq 1/k$ is less than $1/2^m$. On the complement $E(m)^c$ of this set $\forall k \geq m \, |f_{n_k} - f| < 1/k$ holds. This implies that f_n converges uniformly to f on $E(m)^c$.

The conclusion is the following. Consider a sequence f_n that converges to f in the sense of the L^p norm for some p with $1 \leq p < \infty$. Then f_n converges to f in measure. Furthermore, a subsequence converges almost uniformly.

In the following we shall look also at weak L^p convergence with $1 \leq p < \infty$. Let $1/p + 1/q = 1$, so that $1 < q \leq \infty$. The sequence f_n converges to f in the weak L^p sense if for every g in L^q the integrals $\int f_n(x)g(x)\,dx$ converge to $\int f(x)g(x)\,dx$. (When $1 < p < \infty$ weak convergence is the same as weak $*$ convergence.) If f_n converges to f in the weak L^p sense, then there is no guarantee that a subsequence converges any of these pointwise senses. Weak convergence thus does not give any automatic control over nonlinear operations.

Example: Consider the sequence $\sin(n\theta)$ in $L^2([0,\pi])$ for $n = 1, 2, 3, \ldots$. This converges weakly to zero, but there is no subsequence that converges in measure. Also, it behaves very badly under nonlinear operators. For example, the square $\sin^2(n\theta) = (1 - \cos(2n\theta))/2$ converges weakly to $1/2$.

Chapter 14

Weak $*$ compactness

14.1 Weak $*$ compactness in the standard Banach spaces

This chapter treats another kind of compactness. If X is a Banach space, then the space of all continuous linear functionals on X is the dual space X^*. The weak $*$ topology on X^* is the topology of pointwise convergence of these continuous linear functionals. The fundamental theorem is that the unit ball of X^* is compact.

Say that $X = L^p$ with $1 \leq p < \infty$. The the dual space X^* of L^p is L^q, where $1/p + 1/q = 1$. This says that every continuous linear functional on L^p is of the form $f \mapsto \int f(x)g(x)\,dx$ with some g in L^q. Note that $1 < q \leq \infty$.

Take $1 < q \leq \infty$, so that L^q is one of the dual spaces. The weak $*$ topology on L^q is the coarsest topology in which the functionals $g \mapsto \int f(x)g(x)\,dx$ are continuous. The fundamental theorem says that the unit ball $\|g\|_q \leq 1$ is weak $*$ compact. [Warning: When $1 < q < \infty$ the weak $*$ topology is also called the weak topology, for reasons to be explained below.]

What happens when $X = L^\infty$. In that case the dual space X^* is a space of additive set functions with finite total variation. The idea is that for each subset S there is an indicator function 1_S. The value of the set function μ on the set S is the value of the element μ of X^* on 1_S. The linearity condition ensures that this set function is additive for disjoint sets. But it does not guarantee that it is countably additive. The problem is that monotone convergence does not imply uniform convergence. This dual space contains so many objects that it is usually regarded as rather intractable.

It seems more useful to take X as the space C_0 of continuous functions that vanish at infinity. The dual space of this is the space $\mathcal{M}$ of all signed measures with finite total variation. By a miracle, these measures all extend to set functions that are not only additive, but countably additive. (The miracle is Dini's theorem, which says that monotone convergence of continuous functions implies uniform convergence on compact sets. This is applied to continuous functions on the one point compactification.) The weak $*$ topology on this space of measure is the coarsest topology such that all functionals $\mu \mapsto \int f \, dmu$ are continuous. The fundamental theorem says that the space of measures of total variation bounded by one is compact. [Warning: The weak $*$ topology on the space of measures goes by other names, for instance, it is sometimes called the vague topology.]

If g is an L^1 functions, then f times Lebesgue measure defines a signed measure in $\mathcal{M}$. So the natural compactness theorem takes place in the space of signed measures. In fact, it is easy to find a sequence of L^1 functions with norm bounded by one that converge in the weak $*$ sense to a measure. It suffices to look at the functions in a family of approximate delta functions. These converges as measures in the weak $*$ sense to a point measure.

Contrast this with the behavior of an approximate delta function in L^q with $q > 1$. If the integral of $\delta_\epsilon(x) = \delta_1(x/\epsilon)/\epsilon^n$ is one, then the L^q norm is proportional to $\epsilon^{-\frac{n}{p}}$, which is unbounded as $\epsilon \to 0$. If we normalize the functions to have constant L^q norm, then we get a sequence that converges in the weak $*$ sense to zero.

Measures are not functions, and they behave very badly under nonlinear operations. Conclusion: For nonlinear problems involving a weak $*$ compactness argument, it is best to stay away from the space L^1. The other spaces L^q with $1 < q \leq \infty$ have better weak $*$ compactness properties.

14.2 Compactness and minimization

Recall that a topological space is a set X together with a collection of open subsets, closed under unions and finite intersections.

The closed subsets are the complements of the open subsets. Thus a topological space could just as well be defined as a collection of closed subsets, closed under intersections and finite unions.

If $F : X \to Y$ is a function from one topological space to another, then F is said to be continuous if the inverse image of each open subset of Y is an open subset of X.

It is a standard fact that F from X to Y is continuous if and only if the inverse image of every closed set is closed.

Another equivalent formulation is in terms of closures. A function F from X to Y is continuous if and only if for all subsets A of X and points x in the closure of A, the point $f(x)$ is in the closure of the image $F[A]$ of A under F.

A net in X is a function s from a directed set I to X. A net s converges to an element x in X if for every open set U with $x \in U$ there is an element j in the directed set such that s_i is in U for all $i \geq j$. If the directed set I consists of the numbers $\{1, 2, 3, \ldots\}$ then the net is a sequence, and this is the usual definition of convergence of sequences. However the concept of net is more general.

The preceding notions have versions in the language of nets. A point x is in the closure of a set A if and only there is a net s with values in A that converges to x. A function F from X to Y is continuous if and only if whenever s is a net that converges to x the corresponding net $f(s)$ converges to $f(x)$. If the spaces are first countable, then the same results are true with nets replaced by sequences.

A topological space S is said to be compact if every open cover of S has a finite subcover.

A topological space S is compact if and only if every collection of closed subsets of S with empty intersection has a finite subcollection with empty intersection.

A collection of subsets has the finite intersection property if every finite subcollection has non-empty intersection. Thus a space S is compact if and only every collection of closed subsets of S that has the finite intersection property has non-empty intersection. This last statement makes clear that compactness may be thought of as an existence claim.

This again has a formulation in terms of closures. A space S is compact if and only if for every collection of subsets of S with the finite intersection property there is a point in S that is in the closure of each subset in the collection.

There is also a formulation in terms of nets. A space S is compact if and only if every net with range in S has a subnet that converges to a point in S. If the space is first countable, then compactness has a similar characterization in terms of sequences.

Suppose that S is a subset of a topological space X and that S has the induced topology. If X is compact and S closed, then S is compact. On the other hand, if X is a Hausdorff space (in particular if X is a metric space), then if S is compact, S is closed.

Theorem 14.1 *If S is a compact subset of X and F is continuous from X to Y, then the image of S under F is compact.*

14.3 The lower topology

We want to apply the theorem on the image of a compact set when F is a function from X to the interval $(-\infty, +\infty]$ of extended real numbers. We know what this means when we take the usual topology on the interval of extended real numbers.

Consider, however, the following unusual topology on $(-\infty, +\infty]$. The topology is the lower topology, in which the non-trivial open sets are all intervals of the form $(a, +\infty]$. The non-trivial closed sets are thus the intervals $(-\infty, a]$.

This topology is not Hausdorff, but it is first countable. So it is possible to characterize properties of this topology in terms of convergence of sequences. A sequence s converges to a number y if and only if for every $\epsilon > 0$ there is an N such that for all $n \geq N$ we have $s_n > y - \epsilon$. Notice that if a sequence converges to y, then it converges also to every number less than y.

The condition that the sequence s converges to y is equivalent to saying that for every $\epsilon > 0$ there is an N such that the infimum of the s_n for $n \geq N$ exceeds $y - \epsilon$. This just says that $y \leq \liminf s$.

Lemma 14.1 *A non-empty subset A of $(-\infty, +\infty]$ is compact in the lower topology if and only if it contains a minimal element.*

Proof: Suppose that A is compact and non-empty. Consider the collection of all closed sets $(-\infty, b]$ that intersect A. Since these are nested, they have the finite intersection property. Therefore there is a number a that belongs to all these closed sets. This number is a lower bound for a. Furthermore, it belongs to A.

To prove the converse, suppose that the set A does not contain a minimal element. Consider the infimum a of the set, and consider a sequence of closed sets of the form $(-\infty, b]$ with b decreasing to a. Then this sequence has the finite intersection property, but its intersection is empty.

Let F be a function from X to $(-\infty, +\infty]$ with the lower topology. Then F is lower semicontinuous (LSC) if and only for each real a the set of all x such that $a < f(x)$ is open in X. This is the same as saying that for each real a the set of all x such that $f(x) \leq a$ is closed in X.

There is another formulation in terms of closures. Thus F is LSC if and only if for every set A, if x is in the closure of A, then $f(x) \leq \sup f[A]$.

If the topology on X is first countable, then the condition that F is lower semicontinuous is that whenever a sequence s converges to y, then the sequence $F(s)$ satisfies $F(y) \leq \liminf F(s)$.

If F is continuous when $(-\infty, +\infty]$ has its metric topology, then F is continuous when $(-\infty, +\infty]$ has the lower topology. In brief, if F is continuous, then F is LSC.

Theorem 14.2 *Let S be a compact topological space. Let $F : S \to (-\infty, \infty]$ be LSC. If S is non-empty, then there is a point x in S at which F assumes its minimum.*

Proof: It may be worth giving the proof explicitly. Say that F is LSC on the non-empty compact set S. Consider the set of all numbers b for which there is an x in S with $F(x) \leq b$. This set is non-empty. For each b in this set, consider the set of all x in S for which $F(x) \leq b$. Since F is LSC, each such set is closed. These sets have the finite intersection property. Since S is compact, there is an element that belongs to all of these sets. This is the desired element.

14.4 Comparison of topologies

In the above discussion we have considered two possible topologies on the interval $(-\infty, +\infty]$. In the following we need also to consider two topologies on X. The coarser one will be called the weak topology, the finer one will be called the strong topology.

Every open set in the weak topology is an open set in the strong topology. Every closed set in the weak topology is a closed set in the strong topology. Therefore the injection of X with the strong topology into X with the weak topology is continuous.

However every compact set in the strong topology is also compact in the weak topology. Thus the advantage of the weak topology is that it may have more compact sets.

Thus we have four kinds of continuity. We have strong continuity, strong lower semicontinuity, weak continuity, and weak lower semicontinuity. What are the relations between these?

If F is weakly continuous, then F is strongly continuous and also F is weakly lower semicontinuous.

If F is strongly continuous, then F is strongly lower semicontinuous.

If F is weakly lower semicontinuous, then F is strongly lower semicontinuous.

Since compactness is desirable, in the following we shall mainly be concerned with the weak topology on X. In this case we can perhaps hope to prove that a function F is weakly lower semicontinuous. That will be enough to show that it assumes a minimum on every weakly compact subset S.

14.5 Weak $*$ topology

Let X be a Banach space. Of course it always has a norm topology. The weak topology of X is the coarsest topology such that all continuous linear functionals on X given by elements of the dual space X^* are continuous. This topology is the one that is most natural in connection with ideas of the Hahn-Banach theorem and convexity.

The open sets in the weak topology are generated by sets of the form $\{g \mid |\langle u, g-f\rangle| < \epsilon\}$ for f in X and u in X^* and $\epsilon > 0$. These are slabs bounded in one direction and unrestricted in all other directions. The open sets in the norm topology are generated by sets of the form $\{g \mid \|g-f\| < \epsilon\}$. These sets are restricted in all directions. The weak topology is coarser than the norm topology.

Since the intersection of a finite collection of open sets is open, we can think of a typical weak open set as a slab bounded in finitely directions, and unrestricted in all other directions. So it is all too easy to approximate an element f by g in such an open neighborhood. A component of g in the restricted directions must be close to f, but the components of g in other directions are quite free to wander.

A net s with values in X converges to x in the weak topology if and only if the net $\langle u, s\rangle$ converges to $\langle u, x\rangle$ for each u in X^*. It is thus clear that a net s that converges to X in the norm topology also converges to x in the weak topology.

Let X be a Banach space, and let X^* be its dual space. The weak $*$ topology on X^* is the coarsest topology such that all linear functions on X^* given by elements of X are continuous. This is a topology of pointwise convergence of functions. We shall see that it is the topology that is most relevant to compactness.

The open sets in the weak $*$ topology are generated by sets of the form $\{v \mid |\langle v-u, f\rangle| < \epsilon\}$ for u in X^* and f in X and $\epsilon > 0$. The open sets in the norm topology are generated by sets of the form $\{v \mid \|v - u\| < \epsilon\}$, where the norm on X^* is defined by the supremum over the unit ball in X. The weak $*$ topology is coarser than the norm topology.

Again the typical weak $*$ open set is a slab that is restricted in finitely many dimensions. However the directions in which the restriction takes place may be somewhat more limited, since they are given only by those special linear functionals on X^* that come from evaluations at points of X.

A net s with values in X^* converges to u in the weak $*$ topology if and only if the net $\langle s, f\rangle$ converges to $\langle u, f\rangle$ for each f in X. It is thus clear that a net s that converges to X in the norm topology also converges to x in the weak topology.

The space X^* has both a weak topology and a weak $*$ topology. Every element of X is also an element of X^{**}. So every weak $*$ neighborhood is a weak neighborhood. Hence the weak $*$ topology is even coarser than the weak topology. Weak convergence of a net implies weak $*$ convergence.

Consider the weak topology and the weak $*$ topology on X^*. There can be fewer closed sets in the weak $*$ topology. But there can be more compact sets in the weak $*$ topology.

The following is a fundamental result on compactness. Alaoglu: Every closed ball B in X^* is weak $*$ compact.

Here is a corollary. Recall that a closed subset of a compact space is compact. Thus: If a subset S of X^* is weak $*$ closed and bounded in the norm, then it is weak $*$ compact. [Dunford and Schwartz I ; V.4.3] (The converse is also true.)

For each Banach space there is a natural injection of X into X^{**}. A Banach space is reflexive if this is an isomorphism. A reflexive Banach space may be regarded as the dual of its dual. For a reflexive Banach space the weak topology and the weak $*$ topology are the same.

Example: The dual of the space spaces L^p for $1 < p < \infty$ are the spaces L^q for $1 < q < \infty$. The relation between p and q is

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{14.1}$$

Thus these spaces are reflexive.

Example: The dual of L^1 is L^∞. However the dual of L^∞ is typically larger than L^1. Typically L^1 is not a dual space, so there is no weak $*$ topology. The unit ball in L^1 is not weakly compact. One can take an approximate delta function sequence, and it has no subsequence that converges weakly to an element of L^1. As an element of the dual of L^∞ it does converge in the weak $*$ sense to some functional on L^∞, but this functional is not given by a function in L^1.

Why does this sort of example not work for the spaces L^p for $p > 1$? The reason is that if we take a sequence of functions f_k with fixed L^p norm and support on small sets A_k, then $\|f_k\|_1 \leq \mathrm{meas}(A_k)^{\frac{1}{q}}\|f_k\|_p$ approaches zero, since $q < \infty$. So one is getting a delta function, but multiplied by zero.

The fundamental theorem on the existence of a minimum may be stated in this context.

Theorem 14.3 *Let S be a weak $*$ compact subset of X^*. Let F be a function on S that is weak $*$ LSC. If S is non-empty, then there is a point in S at which F assumes its minimum.*

Corollary 14.1 *Let S be a weak $*$ closed subset of X^*. (For example S could be the entire Banach space.) Let F be a function on S that is weak $*$ LSC. Assume that F satisfies the coercivity condition that*

$$\lim_{\|x\|\to\infty} F(x) = +\infty. \tag{14.2}$$

Then there is a point in S on which F assumes its minimum.

Proof: We may suppose that there is an x_1 in S such that $F(x_1) < \infty$. Then there exists k so that if $\|x\| > k$, then $F(x) > F(x_1)$. Let S_k be S intersected with $B(0, k)$. Then S_k is weak $*$ compact. Furthermore, x_1 is in S_k, so S_k is non-empty. Therefore by the theorem, the restriction of F to S_k assumes its minimum at some point x_0. If x is in S, then either x is in S_k, so $F(x_0) \leq F(x)$, or $\|x\| > k$, so $F(x_0) \leq F(x_1) < F(x)$. So x_0 is also a minimum point for the original F.

Technical note: All that is used in the proof is that F is weak $*$ LSC on S intersected with a sufficiently large ball. We shall see that when X is separable the ball is weak $*$ metrizable. So it is sufficient to use sequences to establish this condition.

14.6 Metrizability

Even for a reflexive Banach space the weak topology need not be metrizable. This is not even true for Hilbert space. Here is a simple proof, taken from Halmos (A Hilbert Space Problem Book).

Consider an infinite dimensional Hilbert space with orthonormal basis e_n, $n = 1, 2, 3, \ldots$. Consider the vectors $\sqrt{n}e_n$. First we show that 0 is in the weak closure of this set of vectors.

Let $\{h \mid |\langle h, g_i\rangle| < \epsilon, i = 1, \ldots, k\}$ be a neighborhood of 0. Let the nth Fourier coefficient of g_i be

$$a_{i,n} = \langle e_n, g_i\rangle. \tag{14.3}$$

It is clear from the triangle inequality that

$$\sqrt{\sum_{n=1}^{\infty}(\sum_{i=1}^{k}|a_{i,n}|)^2} \leq \sum_{i=1}^{k}\sqrt{\sum_{n=1}^{\infty}|a_{i,n}|^2} = \sum_{i=1}^{k}\|g_i\| < \infty. \tag{14.4}$$

Therefore it is impossible that for each n

$$\sum_{i=1}^{k}|a_{i,n}| \geq \frac{\epsilon}{\sqrt{n}}. \tag{14.5}$$

As a consequence, for some n

$$\sum_{i=1}^{k} |a_{i,n}| < \frac{\epsilon}{\sqrt{n}}. \tag{14.6}$$

In particular, there exists n such that for each $i = 1, \ldots, k$

$$|a_{i,n}| < \frac{\epsilon}{\sqrt{n}}. \tag{14.7}$$

This says that

$$|\langle \sqrt{n} e_n, g_i \rangle| < \epsilon. \tag{14.8}$$

This shows that this $\sqrt{n}e_n$ is in the weak neighborhood.

We know that there is a net with values in the set of $\sqrt{n}e_n$ that converges to zero. Suppose that the weak topology were metrizable. Then there would be a subsequence f_j of the $\sqrt{n}e_n$ that converges weakly to zero. Thus $\langle f_j, g\rangle$ converges to zero for each g. However a convergent sequence of numbers is bounded. (This statement does not generalize to nets). In particular, the numbers $\langle f_j, g\rangle$ are bounded for each g. However then by the principle of uniform boundedness the f_j are bounded. This is a contradition. We conclude that the weak topology is not metrizable.

It may be shown [Dunford and Schwartz I; V.5.1] that a bounded set in X^* is weak $*$ metrizable if and only if X is a separable metric space. Since most practical applications of Banach spaces use only separable spaces, it follows that we may often use the usual ideas of convergence of sequences in this context. All we need to do is to restrict attention to a bounded set in the Banach space.

The weak $*$ topology on a bounded set may be characterized in various convenient ways. Consider a bounded set B in X^*. Let E be a dense subset of X. The open subsets of B in the weak $*$ topology are generated by sets of the form $\{v \mid |\langle v-u, f\rangle| < \epsilon\}$ for u in X^* and f in E and $\epsilon > 0$. This is a standard 3ϵ argument. Thus weak $*$ convergence on bounded sets is defined by convergence on the set E. For instance, one could take E to be a set of smooth functions with compact support, so that this is convergence in the sense of distributions. Or one could take E to be a set of finite linear combinations of indicator functions of sets of finite measure, so that convergence means convergence of averages over these sets.

14.7 Convexity

The main task is to find a way of proving that a function f is weak $*$ continuous. It turns out that convexity is a key idea.

The following is a fundamental result on closed sets. Mazur: Consider a convex subset S of the Banach space X. If S is closed in the norm, then S is weakly closed. [Dunford and Schwartz I ; V.3.13] (The converse is also true.)

It follows for a reflexive Banach space that a norm closed and bounded convex subset is weak $*$ compact. For example, the unit ball is weak $*$ compact. However the unit sphere is not convex, so it is typically not weak $*$ compact.

Example: The space L^∞ has both a weak topology and a weak $*$ topology. For L^∞ a subset S can be convex and closed, but not weak $*$ closed. For example, fix a point, and take the set S_0 of all functions for which there exist a neighborhood of the point where the function is one. This is a convex set. Every function in this set is a distance at least one from the zero function. Take the norm closure S of this set. This is again a convex set. Again every function in this closed convex set has distance one from the zero function. On the other hand, there is a sequence of functions in the set that converge to zero in the weak $*$ sense. Take g_k to be one on a set of measure $1/k$, zero elsewhere. Then $\int g_k f\, dx \to 0$ for each f in L^1, by the dominated convergence theorem. This proves that the zero function is in the weak $*$ closure of S, but not in the set S itself.

Note that the zero function is not in the weak closure of S. The weak dual of L^∞ might contain a delta measure at the point, and the sequence of values of g_k at the point would not converge.

Why does this sort of example not work for the L^p spaces for $p < \infty$?. The reason is that the set S would already contain the zero function.

Theorem 14.4 *Let S be a norm closed convex subset of a reflexive Banach space X. (Thus S could be all of X.) Let F be a function on X that is convex and norm LSC. Then F is LSC with respect to the weak $*$ topology.*

Proof: We must show that for each a the inverse image of $(-\infty, a]$ under F is weak $*$ closed. However since F is norm LSC, this set is norm closed. Furthermore, since F is convex, the set is also convex. Therefore the set is weakly closed. Since X is reflexive, the set is weak $*$ closed.

Example: Let X be a Hilbert space and let L be a continuous linear functional on X. Let $F(x) = (1/2)\|x\|^2 - L(x)$. Then F is convex and norm continuous. We conclude that it is weak $*$ LSC. Since it also satisfies the coercivity condition, it assumes its minimum at some point z. At this point $\langle z, x\rangle = L(x)$. This gives another approach to the Riesz representation theorem.

Chapter 15

Variational methods for nonlinear problems

15.1 The Euler-Lagrange equation

Consider a smooth function $L(p, z, x)$ defined for vector p, scalar z, and x in some open set U. This is called the Lagrangian. The problem is to minimize the action (or energy) functional

$$I(w) = \int_U L(Dw, w, x)\, dx \tag{15.1}$$

over all functions w defined on U satisfying a boundary condition, say $w = g$ on ∂U.

Suppose that there is a minimum u. Then for each smooth v with compact support in the interior of U we have

$$I(u + v) \geq I(u). \tag{15.2}$$

However

$$I(u + v) = \int_U L(Du + Dv, u + v, x)\, dx \tag{15.3}$$

Thus for small v

$$I(u + v) = I(u) + \int_U [D_p L(Du, u, x) \cdot Dv + D_z L(Du, u, x) v]\, dx + \cdots. \tag{15.4}$$

By integrating by parts, we see that

$$I(u + v) = I(u) + \int_U [-\operatorname{div} D_p L(Du, u, x) + D_z L(Du, u, x)] v\, dx + \cdots. \tag{15.5}$$

The only way that this can happen for all v is that

$$-\operatorname{div} D_p L(Du, u, x) + D_z L(Du, u, x) = 0. \tag{15.6}$$

This is the famous Euler-Lagrange equation.

The conclusion is that minimizing the action should be a way of producing a solution of the Euler-Lagrange equation.

Example: Let $L(p, z, x) = G(p, x) + F(z, x)$. Let $g(p, x) = D_p G(p, x)$ and $f(z, x) = \partial F(z, x)/\partial z$ be the corresponding gradient vector field and scalar derivative. Then the Euler-Lagrange equation says that

$$-\operatorname{div} g(Du, x) + f(u, x) = 0. \tag{15.7}$$

If we define the current $J(Du, x) = -g(Du, x)$ then this is a conservation law

$$\operatorname{div} J(Du, x) = -f(u, x). \tag{15.8}$$

The new features are that the source $-f(u, x)$ is a nonlinear function of the solution u, and the current $J(Du, x)$ is related to the gradient Du by a nonlinear function.

One important special case is when $G(p) = H(|p|^2/2)$. Then the nonlinear vector field $g(p) = h(|p|^2/2)p$ with $h = H'$, and so $g(Du, x) = h(|Du|^2/2)Du$ is a scalar multiple of the usual gradient, where the multiple depends on the magnitude of the gradient.

In the following we treat the problem of minimization directly. However we should recognize that at the same time we are solving a partial differential equation, at least in some weak sense.

15.2 Coercivity

The basic technique is the following. Let X^* be a Banach space that is a dual space. Let I be a function on X^* that is weak $*$ LSC. Assume that I satisfies the coercivity condition that

$$\lim_{\|w\| \to \infty} I(w) = +\infty. \tag{15.9}$$

Then there is a point in X^* on which I assumes its minimum.

The condition that I is weak $*$ LSC says that for every a the set of all w such that $I(w) \leq a$ is weak $*$ closed. The role of the coercivity condition is to ensure that for some k large enough the minimum is assumed on the ball $\|w\| \leq k$, which is compact. Thus it is enough to show that I is weak $*$ LSC on each such ball.

Suppose that X is a separable metric space. Then the weak $*$ topology on such a ball in X^* is metrizable. So it is enough to use sequential convergence to check that I is weak $*$ LSC on each such ball in X^*.

In the application X^* is a space $W_0^{1,q}(U)$, where U is a bounded open set. We always take $1 < q < \infty$. Typically we would like to take q rather small, say close to 1, so that we can get more general result. Sometimes the case $q = 2$ is convenient.

Here are some facts about this Sobolev space:

1. Embedding theorem: $W_0^{1,q}(U) \subset L^q(U)$, and the embedding is compact.
2. Poincaré inequality: $\|u\|_q^q \leq C\|Du\|_q^q$.

Let the Lagrangian function be a smooth function. The function I will be taken to be

$$I(w) = \int_U L(Dw, w, x)\, dx. \tag{15.10}$$

Theorem 15.1 *Suppose that there is an $\alpha > 0$ such that L satisfies the inequality $L(p, z, x) \geq \alpha|p|^q$. Let U be a bounded open set. Then the corresponding functional I satisfies the coercivity condition on $W_0^{1,q}(U)$.*

Proof: The inequality for the Lagrangian implies that $\alpha\|Dw\|_q^q \leq I(w)$. It follows from the Poincaré inequality that the Sobolev norm satisfies $\|w\|_{W_0^{1,q}}^q \leq (C+1)\|Dw\|_q^q$. It follows that the coercivity condition is satisfied.

15.3 The weak $*$ topology on the Sobolev space

.

Let U be an open set. Fix q with $1 < q < \infty$. Let $W_0^{1,q}(U)$ be the Sobolev space of all functions u on U such that u is in $L^q(U)$ and the components of Du are in $L^q(U)$ and that satisfy Dirichlet boundary conditions. The norm is given by

$$\|u\|_{W_0^{1,q}}^q = \|u\|_q^q + \|Du\|_q^q = \|u\|_q^q + \sum_{j=1}^n \|D_j u\|_q^q. \tag{15.11}$$

It would be nice to identify the dual space of $W_0^{1,q}(U)$. Let $1/p + 1/q = 1$. Let f be a function in $L^p(U)$, and let g be a vector field whose components are in $L^p(U)$. The linear functional

$$L(u) = \langle f, u\rangle + \langle g, Du\rangle \tag{15.12}$$

is a linear functional on $W_0^{1,q}(U)$. Furthermore, it is continuous, by the Hölder inequality. So L is an element of the dual space.

The space $W_0^{1,p}(U)$ is a reflexive Banach space, so the weak topology and the weak $*$ topology coincide. By definition of weak convergence, if u_n converges to u weakly, then $L(u_n)$ converges to $L(u)$.

It may be shown that every element of the dual space is given by a functional L of this form. Furthermore, it may be shown that the linear functional L is a Schwartz distribution. See the book on Sobolev spaces by Robert Adams for more information.

Unfortunately, this result does not identify the dual space quite as explicitly as one would like. The reason is that the functional L does not define the pair f, g uniquely. However there is a pair f, g of minimal norm that represents the functional in this way.

To see this, recall that if X is a Banach space and h is an element of X, then there is a non-zero element h^* of the dual space of X such that $\langle h^*, h\rangle = \|h^*\|\|h\|$.

In the case when h is a function belonging to the Banach space L^p, the corresponding element h^* in L^q is given by

$$h^* = |h|^{p-1}\frac{h}{|h|} = |h|^{\frac{p}{q}}\frac{h}{|h|}. \tag{15.13}$$

Note that this is a nonlinear operation except when $p = 2$.

The condition that f, g in L^p are a pair of minimal norm defining an element of the dual space of the Sobolev space is that the corresponding pair f^*, g^* in L^q satisfy $Df^* = g^*$. When $p \neq 2$ this is a somewhat awkward nonlinear condition.

15.4 Convex functionals with a derivative bound

The only remaining thing to check is weak $*$ lower semicontinuity on a compact set. We begin with a simple version with very strong hypotheses.

Theorem 15.2 *Suppose that the Lagrangian $L(p, z, x) \geq 0$ and is convex in p and z for each fixed x. Furthermore, suppose that it satisfies bounds $|D_pL(p, z, x)| \leq C + A|p|^{q-1} + B|z|^{q-1}$ and $|D_zL(p, z, x)| \leq C + A|p|^{q-1} + B|z|^{q-1}$. Then for each k the corresponding functional I is weak $*$ lower semicontinuous on the ball of radius k in $W^{1,q}(U)$.*

Proof: We consider the set of w in the ball of radius k in $W^{1,q}(U)$ such that $I(w) \leq a$. We need to show that this set is weak $*$ closed. Let $u_n \to u$ in the weak $*$ sense of $W_0^{1,q}(U)$ with $I(u_n) \leq a$. We must show that $I(u) \leq a$.

The key step is the convexity:

$$a \geq I[u_n] \geq I[u] + \int_U [D_pL(Du, u, x)(Du_n - Du) + D_zL(Du, u, x)(u_n - u)]\, dx. \tag{15.14}$$

The goal is to take the limit as u_n converges weak $*$ to u and Du_n converges weak $*$ to Du. The space is $L^q(U)$ which is the dual space of $L^p(U)$. The fixed functions $D_pL(Du, u, x)$ and $D_z(Du, u, x)$ thus need to be in $L^p(U)$.

They are each dominated by a sum involving $|Du|^{q-1}$ and $|u|^{q-1}$. , so it is enough to show that each of these is in L^p. However $1/p+1/q = 1$ and therefore $p(q-1) = q$. Thus the pth powers of $|Du|^{q-1}$ and of $|u|^{q-1}$ are $|Du|^q$ and $|u|^q$. Since u belongs to the Sobolev space, these functions are integrable. Therefore the functions $D_pL(Du, u, x)$ and $D_z(Du, u, x)$ are indeed in L^p.

Therefore we may take the weak $*$ limit to get

$$a \geq I[u]. \tag{15.15}$$

15.5 Convex functionals

It would be nice to get the result without the technical seeming derivative bound. This can be done, at the price of slightly complicating the proof.

Theorem 15.3 *Suppose that the Lagrangian $L(p, z, x) \geq 0$ and is convex in p and z for each fixed x. Then for each k the corresponding functional I is weak $*$ lower semicontinuous on the ball of radius k in $W^{1,q}(U)$.*

Proof: We consider the set of w in the ball of radius k in $W^{1,q}(U)$ such that $I(w) \leq a$. We need to show that this set is weak $*$ closed. Let $u_n \to u$ in the weak $*$ sense of $W_0^{1,q}(U)$ with $I(u_n) \leq a$. We must show that $I(u) \leq a$.

Again we want to use convexity. It would be nice if the fixed functions $D_p L(Du, u, x)$ and $D_z L(Du, u, x)$ were in the dual space L^p, but we do not know that.

Let F_ϵ be the set of all x in U where $|u(x)| + |Du(x)| \leq 1/\epsilon$. Then we use positivity to write:

$$a \geq I[u_n] \geq \int_{F_\epsilon} L(Du_n, u, x)\, dx. \tag{15.16}$$

The convexity now gives

$$a \geq \int_{F_\epsilon} L(Du, u, x)\, dx + \int_{F_\epsilon} [D_p L(Du, u, x)(Du_n - Du) + D_z L(Du, u, x)(u_n - u)]\, dx. \tag{15.17}$$

Now the fixed functions are bounded and hence in L^p. So taking the weak $*$ limit we get

$$a \geq \int_{F_\epsilon} L(Du, u, x)\, dx. \tag{15.18}$$

Now let $\epsilon \to 0$ and use the positivity of L and the monotone convergence theorem. This gives

$$a \geq \int_U L(Du, u, x)\, dx = I(u). \tag{15.19}$$

15.6 Functionals convex in the gradient

The preceding theorem uses the convexity very heavily. It would be nice to get rid of it, and the following theorem shows that it is possibility to get a result with only convexity in the derivative. So it is a more powerful and useful result.

Theorem 15.4 *Suppose that the Lagrangian $L(p, z, x) \geq 0$ and is convex in p for each fixed z and x. Then for each k the corresponding functional I is weak $*$ lower semicontinuous on the ball of radius k in $W^{1,q}(U)$.*

Proof: We consider the set of w in the ball of radius k in $W^{1,q}(U)$ such that $I(w) \leq a$. We need to show that this set is weak $*$ closed. Let $u_n \to u$ in the weak $*$ sense of $W_0^{1,q}(U)$ with $I(u_n) \leq a$. We must show that $I(u) \leq a$.

By the compactness of the embedding we can assume we are using a subsequence such that u_n converges to u in the norm of $L^q(U)$. Fix $\epsilon > 0$. It follows that there is a subset E_ϵ of U whose complement has measure less than ϵ and so that u_n converges to u uniformly on E_ϵ.

Let F_ϵ be the set of all x in U where $|u(x)| + |Du(x)| \leq 1/\epsilon$. Finally, let $G_\epsilon = E_\epsilon \cap F_\epsilon$.

Then we use positivity and convexity in the gradient to write

$$a \geq I[u_n] \geq \int_{G_\epsilon} L(Du_n, u_n, x)\,dx \geq \int_{G_\epsilon} L(Du, u_n, x)\,dx + \int_{F_\epsilon} D_pL(Du, u_n, x)(Du_n - Du)\,dx. \tag{15.20}$$

There is considerable advantage to working on the set G_ϵ. Now the function $L(Du, u_n, x)$ is bounded and converges uniformly to $L(Du, u, x)$. Furthermore, $D_pL(Du, u_n, x)$ is bounded and converges uniformly to $D_pL(Du, u, x)$. In particular, it converges in L^p.

We now use the fact that if f_n converges to f in norm in the Banach space and if bounded g_n converges to g in the weak $*$ sense in its dual, then $\langle f_n, g_n\rangle$ converges to $\langle f, g\rangle$. So taking the weak $*$ limit we get

$$a \geq \int_{G_\epsilon} L(Du, u, x)\,dx. \tag{15.21}$$

Now let $\epsilon \to 0$ and use the positivity of L and the monotone convergence theorem. This gives

$$a \geq \int_U L(Du, u, x)\,dx = I(u). \tag{15.22}$$

The final result is the following.

Theorem 15.5 *Consider the Sobolev space $W_0^{1,q}(U)$, where U is a bounded open set, with $1 < q < \infty$. Let the Lagrangian function be a smooth function satisfying for some $\alpha > 0$ the inequality $L(p, z, x) \geq \alpha|p|^q$. Suppose furthermore that $L(p, z, x)$ is convex in p. Define the functional I by*

$$I(w) = \int_U L(Dw, w, x)\,dx. \tag{15.23}$$

Then there is a function u in the Sobolev space such that $I(u) \leq I(w)$ for all w in the Sobolev space.

15.7 Functionals without convexity

We examine Lagrangian functions with and without convexity in the gradient variable. We shall see that a lack of convexity produces big trouble.

Consider $L(p, z) = H(p^2/2) + F(z)$. The functional is

$$I(w) = \int_U [H(|Du|^2/2) + F(u)]\,dx. \tag{15.24}$$

The Euler-Lagrange equation is

$$-\operatorname{div}(h(|Du|^2/2)Du) + f(u) = 0. \tag{15.25}$$

This is a conservation law with current $J(Du) = -h(|Du|^2/2)Du$ and source $-f(u)$. If we consider the corresponding time dependent diffusion, this is

$$\frac{\partial u}{\partial t} = \mathrm{div}(h(|Du|^2/2)Du) - f(u). \tag{15.26}$$

The nicest situation is when both $H(p^2/2)$ and $F(z)$ are convex. The condition for convexity of $H(p^2/2)$ is not difficult to analyze. The first derivative is $D_pH(p^2/2) = h(p^2/2)p$. The second derivative is $D_p^2H(p^2/2) = h'(p^2/2)pp^T + h(p^2/2)I$. So if $h' \geq 0$ and $h \geq 0$ the convexity is satisfied. This says that the nonlinear diffusion coefficient $h(|Du|^2/2)$ is positive and increasing as a function of the size of the gradient. This is a stabilizing effect.

The convexity of $F(z)$ says that its derivative $f(z)$ is increasing. Therefore the source term $-f(u)$ is decreasing as a function of the solution u. This is again a stabilizing effect.

However the theorem applies even without the assumption of convexity of $F(z)$. According to the theorem, we may take $F(z)$ to be an arbitrary smooth function that is positive, or at least bounded below. One way to ensure that $F(z)$ is bounded below is to have its derivative $f(z)$ be bounded above away from zero at $+\infty$ and be bounded below away from zero at $-\infty$. This says that the source $-f(u)$ goes from positive to negative as a function of u. But its behavior in between can be quite complicated.

It is illuminating to look at a specific case. Take the example $L(p, z) = \frac{1}{2}p^2 + \frac{1}{4}(z^2 - 1)^2$. The zero order part is not convex, but this is not a problem. When we minimize

$$I(w) = \int_U [\frac{1}{2}|Dw|^2 + \frac{1}{4}(w^2 - 1)^2]\, dx \tag{15.27}$$

with Dirichlet boundary conditions on a large region we get solutions that in the interior are nearly constant with values ±1. The nonconvexity in the source is not a problem as far as existence of a solution is concerned. (Of course it does create non-uniqueness, but that is another matter.)

For the corresponding diffusion the current is $-Du$ and the source is given by $-u(u^2 - 1)$. The diffusion equation is

$$\frac{\partial u}{\partial t} = \triangle u - u(u^2 - 1). \tag{15.28}$$

Large values of u are damped out, but small values are amplified. The diffusion has an overall smoothing effect.

On the other hand, consider $L(p, z) = \frac{1}{4}(p^2 - 1)^2 + \frac{1}{2}z^2$. When we try to minimize

$$I(w) = \int_U [((Dw)^2 - 1)^2 + w^2]\, dx \tag{15.29}$$

with Dirichlet boundary conditions we encounter a problem. This is particularly easy to see in the one dimensional case. We can take a continuous function that is piecewise linear with slope ±1, and this will make the contribution of the first term equal to zero. By making it change slope many times, we can make the second term arbitrarily small. Therefore the infimum of the $I(w)$ for functions in $W_0^{1,4}$ is equal to zero. However there is no function u with $I(u) = 0$.

It is illuminating to see what is happening from the point of weak $*$ convergence. The functions w are getting small, and their derivatives Dw are oscillating more and more. So they are approaching zero in the weak $*$ topology of $W_0^{1,4}$. However $I(0)$ is not zero. This is a failure of weak $*$ lower semicontinuity. Nonlinear operations can interact with the oscillations of weak $*$ convergence in a very unpleasant way. The minimization forces oscillations, and yet the oscillations can vanish in the limit. No record of their presence is retained in the limiting function. The microstructure of the w with small $I(w)$ might be of interest, but it is not recovered by this method.

For this nasty example the current is $-(|Du|^2 - 1)Du$ and the source is $-u$. The diffusion equation is

$$\frac{\partial u}{\partial t} = \mathrm{div}((|Du|^2 - 1)Du) - u. \tag{15.30}$$

The pathological feature is that a small gradient in density will produce a current that pushes the substance up the gradient. This increases the gradient and produces a severe instability.

Chapter 16

Fixed points

16.1 Banach's fixed point theorem

The following result is Banach's fixed point theorem.

Theorem 16.1 *Let X be a Banach space and let $A : X \to X$ be a nonlinear mapping. Suppose that there is a constant $\gamma < 1$ such that for all u and v in X the inequality*

$$\|A[u] - A[v]\| \leq \gamma \|u - v\| \tag{16.1}$$

is satisfied. Then A has a unique fixed point.

The hypothesis says that A satisfies a Lipshitz condition with Lipshitz constant $\gamma < 1$.

Suppose that A has a derivative dA mapping X to the dual space X^* satisfying

$$\frac{d}{dt} A(u + tw) = \langle dA[u + tw], w \rangle. \tag{16.2}$$

Then in particular

$$A[u] - A[v] = \int_0^1 \langle dA[v + t(u - v)], u - v \rangle \, dt. \tag{16.3}$$

So if the derivative satisfies the bound

$$\|dA[w]\| \leq \gamma \tag{16.4}$$

for all w in X, then A satisfies the Lipschitz condition.

16.2 The Schauder-Leray fixed point theorem

The next result is Schauder's fixed point theorem. It is a generalization of Brouwer's fixed point theorem to the infinite dimensional situation.

Theorem 16.2 *Let X be a Banach space. Let K be a subset that is compact, convex, and non-empty. Assume that $A : K \to K$ is continuous. Then A has a fixed point.*

The proof is obtained by using applying Brouwer's fixed point theorem to a finite dimensional approximation and using compactness to pass to the limit. (The proof is given in Evans.)

Here is an example to show that the theorem can fail when K is a bounded convex set that is not compact. Take the unit ball in Hilbert space. Consider a basis e_j with $j = 1, 2, 3, \ldots$. If A is a linear transformation satisfying $A[e_j] = e_{j+1}$, then A has no fixed point.

Here is an example to show that the theorem can fail when K is a compact set that is not convex. Let the Banach space be two dimensional and let K be a circle. Then an example is given by taking A to be a rotation. Of course the full strength of convexity is not really needed, since this is a topological result. Some of the following results make use of this flexibility.

Corollary 16.1 *Let X be a Banach space. Let E be a subset that is closed, convex, and non-empty. Let $A : E \to E$ be continuous. Suppose that $A[E]$ is contained in a compact set. Then A has a fixed point.*

Proof: Let K be the closure of the set of convex combinations of elements of $A[E]$. Then K is compact, convex, and non-empty. Also K is a subset of E. Furthermore A maps K to itself. So the restriction of A to K has a fixed point.

Corollary 16.2 *Let X be a Banach space. Let B be a closed ball in X. Let $A : B \to X$ be continuous. Assume that the image under A of the ball B is contained in a compact subset of X. Furthermore, assume that the image of the sphere ∂B under A is contained in the interior of the ball B. Then A has a fixed point.*

Proof: Let ϕ map each point in X to the nearest point in B. Let $\tilde{A}[u] = \phi[A[u]]$. Then $\tilde{A}$ maps B to itself and is continuous. So $\tilde{A}$ has a fixed point u in B with $\tilde{A}[u] = u$.

Suppose u is in ∂B. Then $\tilde{A}[u]$ is in ∂B, so $A[u]$ cannot be in the interior of B. This contradicts the hypothesis of the theorem. Thus u must be in the interior of B. Hence $\tilde{A}[u]$ must be in the interior of B. Hence $\tilde{A}[u] = A[u]$. It follows that $A[u] = u$.

The most useful theorem of this type is the following Leray-Schauder fixed point theorem. The setting of the theorem is a homotopy between a constant map and the map A of interest. The hypothesis is that there is a bound on the size of any possible fixed point. The conclusion is that A actually has a fixed point. Notice that this theorem gives Schaefer's fixed point theorem (presented in Evans) as a special case.

Theorem 16.3 *Let X be a Banach space. Let B be a closed ball. Let $F : [0, 1] \times X \to X$ be a continuous map such that the image of $[0, 1] \times B$ is contained* *in a compact subset of X. Suppose that*

$$F(0,u) = u_0 \tag{16.5}$$

is a constant map and

$$F(1,u) = A[u] \tag{16.6}$$

is the map of interest. Suppose that for all t with $0 \leq t \leq 1$ every solution of

$$F(t,u) = u \tag{16.7}$$

lies in the interior of B. Then A has a fixed point.

Proof: Without loss of generality we can take the ball B to be centered at the origin and to have radius one.

Let $0 < \epsilon \leq 1$. Let $\tau_\epsilon(u) = (1 - \|u\|)/\epsilon$ for $\|u\| \geq 1 - \epsilon$ and $\tau_\epsilon(u) = 1$ for $\|u\| \leq 1 - \epsilon$. Define G_ϵ on B by

$$G_\epsilon(u) = F(\tau_\epsilon(u), u). \tag{16.8}$$

Then $G_\epsilon(u)$ maps the boundary ∂B to u_0. It follows that G_ϵ has a fixed point u_ϵ. That is,

$$F(\tau_\epsilon(u_\epsilon), u_\epsilon) = u_\epsilon. \tag{16.9}$$

Since F is compact we can choose a sequence of ϵ tending to zero such that u_ϵ and $\tau_\epsilon(u_\epsilon)$ converge to some u and t. Then

$$F(t,u) = u. \tag{16.10}$$

Suppose $t < 1$. Then $\tau_\epsilon(u_\epsilon)$ is bounded away from 1 for small ϵ. Therefore for small ϵ we have $\|u_\epsilon\| \geq 1 - \epsilon$. It follows that $\|u\| = 1$. This contradicts the assumption that all fixed points are in the interior. So $t = 1$. We conclude that $A(u) = F(1,u) = u$.

16.3 Semilinear elliptic PDE

This section presents a simple illustration of the technique. Let U be a torus and consider the equation

$$-\Delta u = f(x,u). \tag{16.11}$$

for x in U. The advantage of using a torus (periodic boundary conditions) is that one does not have to worry about the boundary at all. This is a reaction diffusion equilibrium equation. It describes the concentration of a substance with source $f(x,u)$ depending on space and on the concentration in some complicated nonlinear way.

Let $a_0 < a_1$ Suppose that f satisfies the stability condition that $u < a_0$ implies $f(x,u) > 0$ and $u > a_1$ implies $f(x,u) < 0$. Thus too low a concentration gives a source and too high a concentration gives a sink.

Lemma 16.1 *Under the stability condition every solution u of the reaction-diffusion equilibrium equation satisfies $a_0 \leq u \leq a_1$.*

Proof: This follows from the maximum principle. If u assumes its maximum at x, then $\Delta u(x) \leq 0$. Therefore from the equation $f(x, u(x)) \geq 0$. It follows that $u(x) \leq a_1$. The other case is similar. .

Theorem 16.4 *Under the stability condition there exists a solution of the reaction-diffusion equilibrium equation.*

Proof: Consider the operator

$$A[u] = (-\Delta + 1)^{-1}[f(x, u) + u]. \tag{16.12}$$

The equation under consideration is equivalent to the fixed point equation $u = A[u]$.

Let b be such that $a_0 < b < a_1$. For each t with $0 \leq t \leq 1$ consider the equation

$$-\Delta u = (1 - t)(b - u) + tf(x, u). \tag{16.13}$$

From the same maximum principle argument it follows that every solution of this equation satisfies $a_0 \leq u \leq a_1$.

The equation can also be written

$$(-\Delta + 1)u = (1 - t)b + t(f(x, u) + u). \tag{16.14}$$

This is the equation $u = F(t, u)$ with

$$F(t, u) = (-\Delta + 1)^{-1}[(1 - t)b + t(f(x, u) + u)]. \tag{16.15}$$

All we need to do is to verify the hypotheses of the Leray-Schauder theorem. The Banach space is $X = C(U)$. Consider a closed ball B in this Banach space that contains all the functions with $a_0 \leq u \leq a_1$ in its interior. Since $(-\Delta+1)^{-1}$ is a compact linear operator, the map F sends $[0, 1] \times B$ into a compact subset of X. Furthermore,

$$F(0, u) = (-\Delta + 1)^{-1}b = b \tag{16.16}$$

and

$$F(1, u) = (\Delta + 1)^{-1}[f(x, u) + u)] = A[u]. \tag{16.17}$$

Furthermore, every solution of $F(t, u) = u$ satisfies $a_0 \leq u \leq a_1$ and hence lies in the interior of B. It follows from the theorem that A has a fixed point.

The proof of this theorem illustrates an important point: An a priori bound on the size of a possible solution can be useful in proving that a solution exists.