Introduction

Inequalities lie at the heart of a great deal of mathematics. G.H. Hardy reported Harald Bohr as saying 'all analysts spend half their time hunting through the literature for inequalities which they want to use but cannot prove'. Inequalities provide control, to enable results to be proved. They also impose constraints; for example, Gromov's theorem on the symplectic embedding of a sphere in a cylinder establishes an inequality that says that the radius of the cylinder cannot be too small. Similar inequalities occur elsewhere, for example in theoretical physics, where the uncertainty principle (which is an inequality) and Bell's inequality impose constraints, and, more classically, in thermodynamics, where the second law provides a fundamental inequality concerning entropy.

Thus there are very many important inequalities. This book is not intended to be a compendium of these; instead, it provides an introduction to a selection of inequalities, not including any of those mentioned above. The inequalities that we consider have a common theme; they relate to problems in real analysis, and more particularly to problems in linear analysis. Incidentally, they include many of the inequalities considered in the fascinating and ground-breaking book *Inequalities*, by Hardy, Littlewood and Pólya [HaLP 52], originally published in 1934.

The first intention of this book, then, is to establish fundamental inequalities in this area. But more importantly, its purpose is to put them in context, and to show how useful they are. Although the book is very largely self-contained, it should therefore principally be of interest to analysts, and to those who use analysis seriously.

The book requires little background knowledge, but some such knowledge is very desirable. For a great many inequalities, we begin by considering sums of a finite number of terms, and the arguments that are used here lie at the heart of the matter. But to be of real use, the results must be extended to infinite sequences and infinite sums, and also to functions and integrals. In order to be really useful, we need a theory of measure and integration which includes suitable limit theorems. In a preliminary chapter, we give a brief account of what we need to know; the details will not be needed, at least in the early chapters, but a familiarity with the ideas and results of the theory is a great advantage.

Secondly, it turns out that the sequences and functions that we consider are members of an appropriate vector space, and that their 'size', which is involved in the inequalities that we prove, is described by a norm. We establish basic properties of normed spaces in Chapter 4. Normed spaces are the subject of linear analysis, and, although our account is largely self-contained, it is undoubtedly helpful to have some familiarity with the ideas and results of this subject (such as are developed in books such as *Linear analysis* by Béla Bollobás [Bol 90] or *Introduction to functional analysis* by Taylor and Lay [TaL 80]. In many ways, this book provides a parallel text in linear analysis.

Looked at from this point of view, the book falls naturally into two unequal parts. In Chapters 2 to 13, the main concern is to establish inequalities between sequences and functions lying in appropriate normed spaces. The inequalities frequently reveal themselves in terms of the continuity of certain linear operators, or the size of certain sublinear operators. In linear analysis, however, there is interest in the structure and properties of linear operators themselves, and in particular in their spectral properties, and in the last four chapters we establish some fundamental inequalities for linear operators.

This book journeys into the foothills of linear analysis, and provides a view of high peaks ahead. Important fundamental results are established, but I hope that the reader will find him- or herself hungry for more. There are brief Notes and Remarks at the end of each chapter, which include suggestions for further reading: a partial list, consisting of books and papers that I have enjoyed reading. A more comprehensive guide is given in the monumental *Handbook of the geometry of Banach spaces* [JoL 01,03] which gives an impressive overview of much of modern linear analysis.

The Notes and Remarks also contain a collection of exercises, of a varied nature: some are five-finger exercises, but some establish results that are needed later. Do them!

Linear analysis lies at the heart of many areas of mathematics, including for example partial differential equations, harmonic analysis, complex analysis and probability theory. Each of them is touched on, but only to a small extent; for example, in Chapter 9 we use results from complex analysis to prove the Riesz-Thorin interpolation theorem, but otherwise we seldom use the powerful tools of complex analysis. Each of these areas has its own collection of important and fascinating inequalities, but in each case it would be too big a task to do them justice here.

I have worked hard to remove errors, but undoubtedly some remain. Corrections and further comments can be found on a web-page on my personal home page at `www.dpmms.cam.ac.uk`

1

Measure and integral

1.1 Measure

Many of the inequalities that we shall establish originally concern finite sequences and finite sums. We then extend them to infinite sequences and infinite sums, and to functions and integrals, and it is these more general results that are useful in applications.

Although the applications can be useful in simple settings – concerning the Riemann integral of a continuous function, for example – the extensions are usually made by a limiting process. For this reason we need to work in the more general setting of measure theory, where appropriate limit theorems hold. We give a brief account of what we need to know; the details of the theory will not be needed, although it is hoped that the results that we eventually establish will encourage the reader to master them. If you are not familiar with measure theory, read through this chapter quickly, and then come back to it when you find that the need arises.

Suppose that Ω is a set. A *measure* ascribes a size to some of the subsets of Ω. It turns out that we usually cannot do this in a sensible way for all the subsets of Ω, and have to restrict attention to the *measurable* subsets of Ω. These are the 'good' subsets of Ω, and include all the sets that we meet in practice. The collection of measurable sets has a rich enough structure that we can carry out countable limiting operations.

A *σ-field* Σ is a collection of subsets of a set Ω which satisfies

(i) if (A_i) is a sequence in Σ then $\cup_{i=1}^{\infty} A_i \in \Sigma$, and

(ii) if $A \in \Sigma$ then the complement $\Omega \setminus A \in \Sigma$.

Thus

(iii) if (A_i) is a sequence in Σ then $\cap_{i=1}^{\infty} A_i \in \Sigma$.

The sets in Σ are called *Σ-measurable* sets; if it is clear what Σ is, they are simply called the *measurable sets*.

Here are two constructions that we shall need, which illustrate how the conditions are used. If (A_i) is a sequence in Σ then we define the *upper limit* $\overline{\lim} A_i$ and the *lower limit* $\underline{\lim} A_i$:

$$\overline{\lim} A_i = \cap_{i=1}^{\infty} \left(\cup_{j=i}^{\infty} A_j \right) \quad \text{and} \quad \underline{\lim} A_i = \cup_{i=1}^{\infty} \left(\cap_{j=i}^{\infty} A_j \right).$$

Then $\overline{\lim} A_i$ and $\underline{\lim} A_i$ are in Σ. You should verify that $x \in \overline{\lim} A_i$ if and only if $x \in A_i$ for infinitely many indices i, and that $x \in \underline{\lim} A_i$ if and only if there exists an index i_0 such that $x \in A_i$ for all $i \geq i_0$.

If Ω is the set $\mathbf{N}$ of natural numbers, or the set $\mathbf{Z}$ of integers, or indeed any countable set, then we take Σ to be the collection $P(\Omega)$ of all subsets of Ω. Otherwise, Σ will be a proper subset of $P(\Omega)$. For example, if $\Omega = \mathbf{R}^d$ (where $\mathbf{R}$ denotes the set of real numbers), we consider the collection of *Borel sets*; the sets in the smallest σ-field that contains all the open sets. This includes all the sets that we meet in practice, such as the closed sets, the G_δ sets (countable intersections of open sets), the F_σ sets (countable unions of closed sets), and so on. The Borel σ-field has the fundamental disadvantage that we cannot give a straightforward definition of what a Borel set looks like – this has the consequence that proofs must be indirect, and this gives measure theory its own particular flavour.

Similarly, if (X, d) is a metric space, then the Borel sets of X are sets in the smallest σ-field that contains all the open sets. [Complications can arise unless (X, d) is separable (that is, there is a countable set which is dense in X), and so we shall generally restrict attention to separable metric spaces.]

We now give a size (non-negative, but possibly infinite or zero) to each of the sets in Σ. A *measure* on a σ-field Σ is a mapping μ from Σ into $[0, \infty]$ satisfying

(i) $\mu(\emptyset) = 0$, and

(ii) if (A_i) is a sequence of disjoint sets in Σ then $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$: μ is *countably additive.*

The most important example that we shall consider is the following. There exists a measure λ *(Borel measure)* on the Borel sets of $\mathbf{R}^d$ with the property that if A is the rectangular parallelopiped $\prod_{i=1}^{d}(a_i, b_i)$ then $\lambda(A)$ is the product $\prod_{i=1}^{d}(b_i - a_i)$ of the length of its sides; thus λ gives familiar geometric objects their natural measure. As a second example, if Ω is a countable set, we can define $\#(A)$, or $|A|$, to be the number of points, finite or infinite, in A; $\#$ is *counting measure.* These two examples are radically different: for counting measure, the one-point sets $\{x\}$ are *atoms*; each has positive measure, and any subset of it has either the same measure or zero measure. Borel measure on $\mathbf{R}^d$ is *atom-free*; no subset is an atom. This is equivalent to requiring that if A is a set of non-zero measure A, and if $0 < \beta < \mu(A)$ then there is a measurable subset B of A with $\mu(B) = \beta$.

Countable additivity implies the following important continuity properties:

(iii) if (A_i) is an increasing sequence in Σ then

$$\mu(\cup_{i=1}^{\infty} A_i) = \lim_{i \to \infty} \mu(A_i).$$

[Here and elsewhere, we use 'increasing' in the weak sense: if $i < j$ then $A_i \subseteq A_j$. If $A_i \subset A_j$ for $i < j$, then we say that (A_i) is 'strictly increasing'. Similarly for 'decreasing'.]

(iv) if (A_i) is a decreasing sequence in Σ and $\mu(A_1) < \infty$ then

$$\mu(\cap_{i=1}^{\infty} A_i) = \lim_{i \to \infty} \mu(A_i).$$

The finiteness condition here is necessary and important; for example, if $A_i = [i, \infty) \subseteq \mathbf{R}$, then $\lambda(A_i) = \infty$ for all i, but $\cap_{i=1}^{\infty} A_i = \emptyset$, so that $\lambda(\cap_{i=1}^{\infty} A_i) = 0$.

We also have the following consequences:

(v) if $A \subseteq B$ then $\mu(A) \leq \mu(B)$;

(iv) if (A_i) is any sequence in Σ then $\mu(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$.

There are many circumstances where $\mu(\Omega) < \infty$, so that μ only takes finite values, and many where $\mu(\Omega) = 1$. In this latter case, we can consider μ as a probability, and frequently denote it by $\mathbf{P}$. We then use probabilistic language, and call the elements of Σ 'events'.

A *measure space* is then a triple (Ω, Σ, μ), where Ω is a set, Σ is a σ-field of subsets of Ω (the *measurable sets*) and μ is a measure defined on Σ. In order to avoid tedious complications, we shall restrict our attention to *σ-finite* measure spaces: we shall suppose that there is an increasing sequence (C_k) of measurable sets of finite measure whose union is Ω. For example, if λ is Borel measure then we can take $C_k = \{x \colon |x| \leq k\}$.

Here is a useful result, which we shall need from time to time.

Proposition 1.1.1 (The first Borel–Cantelli lemma) *If (A_i) is a sequence of measurable sets and $\sum_{i=1}^{\infty} \mu(A_i) < \infty$ then $\mu(\overline{\lim} A_i) = 0$.*

Proof For each i, $\mu(\overline{\lim} A_i) \leq \mu(\cup_{j=i}^{\infty} A_j)$, and $\mu(\cup_{j=i}^{\infty} A_j) \leq \sum_{j=i}^{\infty} \mu(A_j) \to 0$ as $i \to \infty$. □

If $\mu(A) = 0$, A is called a *null set*. We shall frequently consider properties which hold except on a null set: if so, we say that the property holds *almost everywhere*, or, in a probabilistic setting, *almost surely*.

1.2 Measurable functions

We next consider functions defined on a measure space (Ω, Σ, μ). A real-valued function f is Σ-*measurable*, or more simply *measurable*, if for each real α the set $(f > \alpha) = \{x \colon f(x) > \alpha\}$ is in Σ. A complex-valued function is *measurable* if its real and imaginary parts are. (When $\mathbf{P}$ is a probability measure and we are thinking probabilistically, a measurable function is called a *random variable*.) In either case, this is equivalent to the set $(f \in U) = \{x \colon f(x) \in U\}$ being in Σ for each open set U. Thus if Σ is the Borel σ-field of a metric space, then the continuous functions are measurable. If f and g are measurable then so are $f + g$ and fg; the measurable functions form an algebra $\mathcal{M} = \mathcal{M}(\Omega, \Sigma, \mu)$. If f is measurable then so is $|f|$. Thus in the real case $\mathcal{M}$ is a lattice: if f and g are measurable, then so are $f \vee g = \max(f, g)$ and $f \wedge g = \min(f, g)$.

We can also consider the Borel σ-field of a compact Hausdorff space (X, τ): but it is frequently more convenient to work with the *Baire σ-field*: this is the smallest σ-field containing the closed G_δ sets, and is the smallest σ-field for which all the continuous real-valued functions are measurable. When (X, τ) is metrizable, the Borel σ-field and the Baire σ-field are the same.

A measurable function f is a *null function* if $\mu(f \neq 0) = 0$. The set $\mathcal{N}$ of null functions is an ideal in $\mathcal{M}$. In practice, we identify functions which are equal almost everywhere: that is, we consider elements of the quotient space $M = \mathcal{M}/\mathcal{N}$. Although these elements are equivalence classes of functions, we shall tacitly work with representatives, and treat the elements of M as if they were functions.

What about the convergence of measurable functions? A fundamental problem that we shall frequently consider is 'When does a sequence of measurable functions converge almost everywhere?' The first Borel–Cantelli lemma provides us with the following useful criterion.

Proposition 1.2.1 *Suppose that (f_n) is a decreasing sequence of non-negative measurable functions. Then $f_n \to 0$ almost everywhere if and only if $\mu((f_n > \epsilon) \cap C_k) \to 0$ as $n \to \infty$ for each k and each $\epsilon > 0$.*

Proof Suppose that (f_n) converges almost everywhere, and that $\epsilon > 0$. Then $((f_n > \epsilon) \cap C_k)$ is a decreasing sequence of sets of finite measure, and if $x \in \cap_n (f_n > \epsilon) \cap C_k$ then $(f_n(x))$ does not converge to 0. Thus, by condition (iv) above, $\mu((f_n > \epsilon) \cap C_k) \to 0$ as $n \to \infty$.

For the converse, we use the first Borel–Cantelli lemma. Suppose that the condition is satisfied. For each n there exists N_n such that $\mu((f_{N_n} > 1/n) \cap C_n) < 1/2^n$. Then since $\sum_{n=1}^{\infty} \mu((f_{N_n} > 1/n) \cap C_n) < \infty$,

$\mu(\overline{\lim}((f_{N_n} > 1/n) \cap C_n) = 0$. But if $x \notin \overline{\lim}((f_{N_n} > 1/n) \cap C_n)$ then $f_n \to 0$. □

Corollary 1.2.1 *A sequence (f_n) of measurable functions converges almost everywhere if and only if*

$$\mu\left((\sup_{m,n \geq N} |f_m - f_n| > \epsilon) \cap C_k \right) \to 0 \quad \text{as } N \to \infty$$

for each k and each $\epsilon > 0$.

It is a straightforward but worthwhile exercise to show that if $f(x) = \lim_{n\to\infty} f_n(x)$ when the limit exists, and $f(x) = 0$ otherwise, then f is measurable.

Convergence almost everywhere cannot in general be characterized in terms of a topology. There is however a closely related form of convergence which can. We say that $f_n \to f$ *locally in measure* (or *in probability*) if $\mu((|f_n - f| > \epsilon) \cap C_k) \to 0$ as $n \to \infty$ for each k and each $\epsilon > 0$; similarly we say that (f_n) is locally *Cauchy in measure* if $\mu((|f_m - f_n| > \epsilon) \cap C_k) \to 0$ as $m, n \to \infty$ for each k and each $\epsilon > 0$. The preceding proposition, and another use of the first Borel–Cantelli lemma, establish the following relations between these ideas.

Proposition 1.2.2 *(i) If (f_n) converges almost everywhere to f, then (f_n) converges locally in measure.*

(ii) If (f_n) is locally Cauchy in measure then there is a subsequence which converges almost everywhere to a measurable function f, and $f_n \to f$ locally in measure.

Proof (i) This follows directly from Corollary 1.2.1.

(ii) For each k there exists N_k such that $\mu((|f_m - f_n| > 1/2^k) \cap C_k) < 1/2^k$ for $m, n > N_k$. We can suppose that the sequence (N_k) is strictly increasing. Let $g_k = f_{N_k}$. Then $\mu((|g_{k+1} - g_k| < 1/2^k) \cap C_k) < 1/2^k$. Thus, by the First Borel–Cantelli Lemma, $\mu(\overline{\lim}((|g_{k+1} - g_k| > 1/2^k) \cap C_k)) = 0$. But $\overline{\lim}(|g_{k+1} - g_k| > 1/2^k) \cap C_k) = \overline{\lim}(|g_{k+1} - g_k| > 1/2^k)$. If $x \notin \overline{\lim}(|g_{k+1} - g_k| > 1/2^k)$ then $\sum_{k=1}^{\infty} |g_{k+1}(x) - g_k(x)| < \infty$, so that $(g_k(x))$ is a Cauchy sequence, and is therefore convergent.

Let $f(x) = \lim g_k(x)$, when this exists, and let $f(x) = 0$ otherwise. Then (g_k) converges to f almost everywhere, and locally in measure. Since $(|f_n - f| > \epsilon) \subseteq (|f_n - g_k| > \epsilon/2) \cup (|g_k - f| > \epsilon/2)$, it follows easily that $f_n \to f$ locally in measure. □

In fact, there is a complete metric on M under which the Cauchy sequences are the sequences which are locally Cauchy in measure, and the convergent sequences are the sequences which are locally convergent in measure. This completeness result is at the heart of very many completeness results for spaces of functions.

If A is a measurable set, its *indicator function* I_A, defined by setting $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ otherwise, is measurable. A *simple function* is a measurable function which takes only finitely many values, and which vanishes outside a set of finite measure: it can be written as $\sum_{i=1}^n \alpha_i I_{A_i}$, where $A_1, \ldots, A_n$ are measurable sets of finite measure (which we may suppose to be disjoint).

Proposition 1.2.3 *A non-negative measurable function f is the pointwise limit of an increasing sequence of simple functions.*

Proof Let $A_{j,n} = (f > j/2^n)$, and let $f_n = \frac{1}{2^n}\sum_{j=1}^{4^n} I_{A_{j,n}\cap C_n}$. Then (f_n) is an increasing sequence of simple functions, which converges pointwise to f. □

This result is extremely important; we shall frequently establish inequalities for simple functions, using arguments that only involve finite sums, and then extend them to a larger class of functions by a suitable limiting argument. This is the case when we consider integration, to which we now turn.

1.3 Integration

Suppose first that $f = \sum_{i=1}^n \alpha_i I_{A_i}$ is a non-negative simple function. It is then natural to define the integral as $\sum_{i=1}^n \alpha_i \mu(A_i)$. It is easy but tedious to check that this is independent of the representation of f. Next suppose that f is a non-negative measurable function. We then define

$$\int_\Omega f\, d\mu = \sup\{\int g\, d\mu \colon g \text{ simple}, 0 \le g \le f\}.$$

A word about notation: we write $\int_\Omega f\, d\mu$ or $\int f\, d\mu$ for brevity, and $\int_\Omega f(x)\, d\mu(x)$ if we want to bring attention to the variable (for example, when f is a function of more than one variable). When integrating with respect to Borel measure on $\mathbf{R}^d$, we shall frequently write $\int_{\mathbf{R}^d} f(x)\, dx$, and use familiar conventions such as $\int_a^b f(x)\, dx$. When $\mathbf{P}$ is a probability measure, we write $\mathbf{E}(f)$ for $\int f\, d\mathbf{P}$, and call $\mathbf{E}(f)$ the *expectation* of f.

We now have the following fundamental continuity result:

Proposition 1.3.1 (The monotone convergence theorem) *If (f_n) is an increasing sequence of non-negative measurable functions which converges pointwise to f, then $(\int f_n\,d\mu)$ is an increasing sequence and $\int f\,d\mu = \lim_{n\to\infty}\int f_n\,d\mu$.*

Corollary 1.3.1 (Fatou's lemma) *If (f_n) is a sequence of non-negative measurable functions then $\int(\liminf f_n)\,d\mu \le \liminf \int f_n\,d\mu$. In particular, if f_n converges almost everywhere to f then $\int f\,d\mu \le \liminf \int f_n\,d\mu$.*

We now turn to functions which are not necessarily non-negative. A measurable function f is *integrable* if $\int f^+\,d\mu < \infty$ and $\int f^-\,d\mu < \infty$, and in this case we set $\int f\,d\mu = \int f^+\,d\mu - \int f^-\,d\mu$. Clearly f is integrable if and only if $\int |f|\,d\mu < \infty$, and then $|\int f\,d\mu| \le \int |f|\,d\mu$. Thus the integral is an absolute integral; fortuitous cancellation is not allowed, so that for example the function $\sin x/x$ is not integrable on $\mathbf{R}$. Incidentally, integration with respect to Borel measure extends proper Riemann integration: if f is Riemann integrable on $[a,b]$ then f is equal almost everywhere to a Borel measurable and integrable function, and the Riemann integral and the Borel integral are equal.

The next result is very important.

Proposition 1.3.2 (The dominated convergence theorem) *If (f_n) is a sequence of measurable functions which converges pointwise to f, and if there is a measurable non-negative function g with $\int g\,d\mu$ such that $|f_n| \le g$ for all n, then $\int f_n\,d\mu \to \int f\,d\mu$ as $n \to \infty$.*

This is a precursor of results which will come later; provided we have some control (in this case provided by the function g) then we have a good convergence result. Compare this with Fatou's lemma, where we have no controlling function, and a weaker conclusion.

Two integrable functions f and g are equal almost everywhere if and only if $\int |f-g|\,d\mu = 0$, so we again identify integrable functions which are equal almost everywhere. We denote the resulting space by $L^1 = L^1(\Omega,\Sigma,\mu)$; as we shall see in Chapter 4, it is a vector space under the usual operations.

Finally, we consider repeated integrals. If (X,Σ,μ) and (Y,T,ν) are measure spaces, we can consider the σ-field $\sigma(\Sigma\times T)$, which is the smallest σ-field containing $A\times B$ for all $A\in\Sigma$, $B\in T$, and can construct the product measure $\mu\times\nu$ on $\sigma(\Sigma\times T)$, with the property that $(\mu\times\nu)(A\times B) = \mu(A)\nu(B)$. Then the fundamental result, usually referred to as *Fubini's theorem*, is that