Introduction

In May 2017, I was privileged to be invited to give a mini-course of lectures at the University of Costa Rica in San José. The participants were mainly a group of advanced undergraduate students who had already taken courses in functional analysis and measure theory. After some discussions with my hosts, we decided that I would lecture on the topic of this book, which has since evolved from the manuscript that I prepared for the lectures into the volume that you are now reading.

Why semigroups? First I should emphasise that I am not an "expert" in this subject, but despite this, it seems that I have, in some sense, spent my entire research career working on this topic, in that whichever problem I happen to be working on, there is always a semigroup either blatantly hogging the limelight in the foreground, or "lurking in the background".[1] The subject is undeniably very attractive from the point of view of intellectual beauty. It is a delightful tour-de-force of fascinating mathematical ideas, which forms a very natural second course in functional analysis, for those who have already had some grounding in the structure of Banach and Hilbert spaces, and associated linear operators. But it is also very deeply connected with applications that both illustrate the theory itself, and also provide the impetus for new theoretical developments. These include, but are not restricted to, partial differential equations, stochastic processes, dynamical systems and quantum theory, and each of these topics features within the current book.

There are already many excellent existing books on semigroup theory and applications, so why publish another? Firstly, I have tried to write a book that exhibits the spirit of my lectures. So I am assuming that the primary readership will consist of final-year undergraduates, MSc students and beginning

[1] Apologies to Nick Bingham for stealing one of his favourite catchphrases.

PhD students, and there is a corresponding pedagogic approach that assumes somewhat less sophistication and experience on behalf of the reader than is usually found in books on this subject. Secondly, I have tried to bring interesting classes of examples into play at an early stage; and this means that there is far more interplay between functional analysis and other areas of analysis – in particular, measure theoretic probability – than in other accounts. Of course, much interdisciplinary research involves the interaction of one or more branches of mathematics, and I hope that readers will benefit by being exposed to this way of thinking within the current text.

The semigroups that we are concerned with will be families of bounded linear operators $(T_t, t \geq 0)$ acting in a real or complex Banach space E, which satisfy the semigroup property $T_{s+t} = T_s T_t$ for all $t \geq 0$, with T_0 being the identity operator on E, and which have nice continuity properties. We will see that there is then a linear operator A acting in E, which is typically not bounded, which is obtained by differentiating the semigroup at $t = 0$. We call A the generator of the semigroup, and one of the key themes of the first two chapters is the interplay between the semigroup and the generator. We do not assume prior knowledge of the theory of unbounded operators, and seek to develop what is needed as we go along. It is natural to interpret the parameter t as describing the flow of time, and the semigroup itself as the dynamical time-evolution of some system of interest. Then the semigroup and its generator represent the global and local descriptions of the dynamics (respectively). Many examples of this type come from partial differential equations (PDEs), where the generator A is typically a second-order (elliptic) differential operator. Other important examples come from the world of stochastic processes, where the semigroup is obtained by averaging over all possible trajectories of some random dynamical evolution. A fascinating aspect of working in this area is the appreciation that these application areas are not distinct, so we can and will approach the same phenomenon from the point of view of semigroup theory, probability theory and PDEs.

The first chapter and first two sections of the second are designed to give a pretty rigorous and thorough introduction to the basic concepts of the theory. In particular, the first part of Chapter 2 presents proofs of the three key theorems that give necessary and sufficient conditions for a linear operator A to be the generator of a semigroup, these being the Feller–Miyadera–Phillips, Hille–Yosida and Lumer–Phillips theorems. Having reached this point, we do not feel overly constrained to give fully mathematically rigorous accounts of what follows. Our goal is more to present a wide range of different topics so that readers can get an introduction to the landscape, but to give only partial proofs where there are too many technical details, or even just heuristic arguments.

In the latter cases, we are of course very careful to give precise references to where detailed proofs can be found.

Continuing our brief tour of the content of the book, the second part of Chapter 2 deals with partial differential equations, and the main point is to show that solutions of second-order parabolic equations can be represented as semigroup actions. Chapter 3 is about semigroups of operators that are obtained from convolution semigroups of measures. To some extent, it is a companion piece to Chapters 1 and 3 of my earlier book [6], but in that work, the emphasis was on the underlying stochastic processes, whereas here we take a more analytic perspective and place the semigroup in centre ground. We present a proof (at least in outline) of the famous Lévy–Khintchine formula which characterises the convolution semigroup through its Fourier transform. The generators are then conveniently represented as pseudo-differential operators, and we also include an introduction to this important theme of modern analysis for those readers who have not met them before.

Three of the most important classes of operators in Hilbert space that are commonly encountered are the self-adjoint, compact and trace class. In Chapter 4, we study self-adjoint semigroups, but also unitary groups, that are generated by iA, where A is self-adjoint. This two-way relationship between skew-adjoint generators and unitary groups, is the content of Stone's theorem, which is proved in this chapter. This paves the way for a discussion of quantum mechanics, as the key Schrödinger equation is an infinitesimal expression of Stone's theorem. Being group dynamics, this is reversible; but we also discuss irreversible dynamics in the quantum context. We do not prove the celebrated Gorini–Kossakowski–Sudarshan–Lindlad theorem that classifies generators in this context, but we do give some probabilistic insight into how such operators can arise. Chapter 5 is concerned with compact and trace class semigroups. We investigate eigenfunction expansions for the semigroup and its kernel (when it acts as an integral operator), and also meet the important Mercer's theorem that relates the trace to the kernel. In both Chapters 4 and 5, the convolution semigroups studied in Chapter 3 are put to work to yield important classes of examples.

In Chapter 6, we take a brief look at perturbation theory, and conclude by giving two derivations of the celebrated Feynman–Kac formula, one based on the Lie–Kato–Trotter product formula, which is proved herein, and the other using Itô calculus. Chapter 7 returns to the theme of Chapter 3, but in greater generality as the context is now Markov and Feller semigroups. Here we give a partial proof of the Hille–Yosida–Ray theorem that gives necessary and sufficient conditions for an operator to generate a positivity-preserving contraction semigroup. One of these conditions is the positive maximum theorem, and we

include a full proof of the Courrège theorem that gives the characteristic form of operators that obey this principle. Our proof uses some ideas from the theory of distributions, and again we give a brief self-contained synopsis of all the material that we'll need.

In Chapter 8, we look more carefully at the relationship between semigroups and (semi)-dynamical systems. The main idea here is that we have a group (say) of transformations on a locally compact space S that expresses some dynamical law. We pull back these transformations to a group of operators acting in a suitable L^2-space. The aim is then to study the original dynamical system from the point of view of the group of operators. This chapter also contains a brief discussion of mathematical models of the origins of irreversibility. Here we try to engage, in a mathematical way, with the fascinating question: is irreversible evolution a fundamental aspect of the way in which Nature works, or just a feature of our inability to see into the real nature of the dynamics, which is reversible?

Finally in Chapter 9, we introduce a class of semigroups on function spaces, called Varopoulos semigroups, after N. Varopoulos who invented them. They are closely related to ultracontractive semigroups. We prove that the Riesz potential operators that occur as the Mellin transforms of Varopoulos semigroups satisfy the Hardy–Littlewood–Sobolev inequality, and obtain some Sobolev inequalities as a special case. We then show that second-order elliptic partial differential operators on bounded regions generate Varopoulos semigroups, and prove the famous Nash inequality along the way.

As can be seen from this synopsis, as well as introducing and developing semigroup theory, our journey enables us to touch on a number of important topics within contemporary analysis. However, in order to keep this introductory book to manageable size, many interesting topics were omitted, such as analytic semigroups,[2] subordination and Dirichlet forms. Of course readers wanting to know more about these and other topics will find ample resources in the bibliography. Where there is more than one edition of a book listed there, I have generally used the more recent one. Each of the first six chapters of the book concludes with a set of exercises. Solutions to these will be made available at www.cambridge.org/9781108483094. The proofs of some theorems that are either only stated in the main text (such as the Lax–Milgram theorem in Chapter 2), or which lie outside the central scope of the book, but are nonetheless both interesting and important (such as the basic criteria for self-adjointness in Chapter 4), appear in the exercises with sufficient guidance for readers to be able to construct

[2] These are in fact given a very short treatment in subsection 6.1.2.

these for themselves. When in the main text, you encounter a reference to Problem x.y, this means the yth problem in the exercises at the end of Chapter x.

It is a pleasure to thank my former PhD student (now a lecturer at the University of Costa Rica) Christian Fonseca Mora, for inviting me there and for working so hard to make my visit enjoyable. I will never forget the warm and generous hospitality extended to me by Christian, his wife Yeime and their extended family. It was also heartening to meet so many keen students, who were exceptionally well-prepared for my lectures, and so hungry to learn modern analysis. I hope that I was able to satisfy their appetites, at least for a short while. Thanks are also due to my PhD students Rosemary Shewell Brockway and Trang Le Ngan, and my colleagues Nic Freeman and Koji Ohkitani, who attended a short informal course based on this material at the University of Sheffield in spring 2018. Last, but not least (on the academic side), I would like to thank both Christian Fonseca Mora and Gergely Bodó (a former undergraduate student at the University of Sheffield) for their careful reading of the manuscript, enabling me to correct many typos and minor errors. Finally it is a pleasure to thank all the hard-working staff at Cambridge University Press who have helped to transform my manuscript into the high-quality book (or e-book) that you are currently reading, particularly the editors Roger Astley and Clare Dennison, project manager Puviarassy Kalieperumal, and copyeditor Bret Workman.

Guide to Notation and a Few Useful Facts

If S is a set, S^c denotes its complement. If T is another set, then $S\backslash T := S\cap T^c$. If A is a finite set, then the number of elements in A is $\#A$. If A is a non-empty subset of a topological space, then $\overline{A}$ is its closure. If S is a metric space, with metric d, then for each $x \in S$, $B_r(x) := \{y \in S; d(x, y) < r\}$ is the open ball of radius $r > 0$, centred at x.

$\mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ are the sets of natural numbers, integers, rational numbers, real numbers and complex numbers (respectively). $\mathbb{Z}_+ := \mathbb{N} \cup \{0\}$. In this short section we use F to denote $\mathbb{R}$ or $\mathbb{C}$.

If S is a topological space, then the *Borel* σ-algebra of S will be denoted $\mathcal{B}(S)$. It is the smallest σ-algebra of subsets of S which contains all the open sets. Sets in $\mathcal{B}(S)$ are called *Borel sets*. $\mathbb{R}$ and $\mathbb{C}$ will always be assumed to be equipped with their Borel σ-algebras, and measurable functions from $(S, \mathcal{B}(S))$ to $(F, \mathcal{B}(F))$ are sometimes called *Borel measurable*. Similarly, a measure defined on $(S, \mathcal{B}(S))$ is called a *Borel measure*.

If S is a locally compact Hausdorff space[3] (we will not meet these general spaces until Chapter 6), then $B_b(S, F)$ is the linear space (with the usual pointwise operations of addition and scalar multiplication) of all bounded Borel measurable functions from S to F. It is an F-Banach space under the supremum norm $||f||_\infty := \sup_{x\in S} |f(x)|$ for $f \in B_b(S, F)$. The space of bounded continuous functions from S to F is denoted $C_b(S, F)$. It is a closed linear subspace of $B_b(S, F)$, and so an F-Banach space in its own right. A function f from S to F is said to *vanish at infinity* if given any $\epsilon > 0$ there exists a compact set K in S so that $|f(x)| < \epsilon$ whenever $x \in K^c$. The space $C_0(S, F)$ of all continuous F-valued functions on S which vanish at infinity[4] is a closed linear subspace of $B_b(S, F)$ (and of $C_b(S, F)$), and so is also an F-Banach space in its own right. The *support* of an F-valued function f defined on S is the closure of the set $\{x \in S; f(x) \neq 0\}$ and it is denoted $\text{supp}(f)$. The linear space $C_c(S, F)$ of all continuous F-valued functions on S with compact support is a dense subspace of $C_0(S, F)$. When $F = \mathbb{R}$, we usually write $B_b(S) := B_b(S, \mathbb{R})$, $C_0(S) := C_0(S, \mathbb{R})$ etc., and this will be our default assumption.

Throughout this book "smooth" means "infinitely differentiable". We write $C(\mathbb{R}^d)$ for the linear space of all continuous real-valued functions on $\mathbb{R}^d$. If $n \in \mathbb{N}$, $C^n(\mathbb{R}^d)$ is the linear space of all n-times real-valued differentiable functions on $\mathbb{R}^d$ that have continuous partial derivatives to all orders. We define $C^\infty(\mathbb{R}^d) := \bigcap_{n\in\mathbb{N}} C^n(\mathbb{R}^d)$, and $C_c^n(\mathbb{R}^d) := C^n(\mathbb{R}^d) \cap C_c(\mathbb{R}^d)$ for all $n \in \mathbb{N} \cup \{\infty\}$.

If $n \in \mathbb{N}$, then $M_n(F)$ is the F-algebra of all $n \times n$ matrices with values in F. The identity matrix in $M_n(F)$ is denoted by I_n. If $A \in M_n(F)$, then A^T denotes its transpose and $A^* = \overline{A^T}$ is its adjoint. The trace of a square matrix A, i.e., the sum of its diagonal entries, is denoted by tr(A), and its determinant is $\det(A)$. The matrix A is said to be *non-negative definite* if $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$, and *positive definite* if $x^T Ax > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$.

If (S, Σ, μ) is a measure space and $f : S \to F$ is an integrable function, we often write the Lebesgue integral $\int_S f(x)\mu(dx)$ as $\mu(f)$. For $1 \leq p < \infty$, $L^p(S) := L^p(S, \mathbb{C}) = L^p(S, \Sigma, \mu; F; \mathbb{C})$ is the usual L^p space of equivalence classes of complex-valued functions that agree almost everywhere with respect to μ for which

$$||f||_p = \left(\int_S |f(x)|^p \mu(dx)\right)^{\frac{1}{p}} < \infty$$

[3] Readers who have not yet learned these topological notions are encouraged to take $S \subseteq \mathbb{R}^d$.
[4] This space will play an important role in this book. Some of its key properties are proved in Appendix A.

for all $f \in L^p(S)$. $L^p(S)$ is a Banach space with respect to the norm $||\cdot||_p$, and $L^2(S)$ is a Hilbert space with respect to the inner product

$$\langle f, g\rangle := \int_S f(x)\overline{g(x)}\mu(dx)$$

for $f, g \in L^2(S)$. The spaces $L^p(S, \Sigma, \mu; F; \mathbb{R})$ are defined similarly. For applications to probability theory and PDEs, we will usually write $L^p(S) := L^p(S, \mathbb{R})$, but if we deal with Fourier transforms or (in the case $p = 2$) quantum mechanics, then we will need complex-valued functions.

The *indicator function* 1_A of $A \in \Sigma$ is defined as follows:

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

If μ is σ-finite, and ν is a finite measure on (S, Σ), we write $\nu \ll \mu$ if ν is absolutely continuous with respect to μ, and $\dfrac{d\nu}{d\mu}$ is the corresponding Radon–Nikodym derivative (see Appendix E).

If $(\Omega, \mathcal{F}, P)$ is a probability space and $X : \Omega \to \mathbb{R}$ is a random variable (i.e., a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$) that is also integrable in that $\int_\Omega |X(\omega)|P(d\omega) < \infty$, then its *expectation* is defined to be

$$\mathbb{E}(X) := \int_\Omega X(\omega)P(d\omega).$$

If $T : V_1 \to V_2$ is a linear mapping between F-vector spaces V_1 and V_2, then $\mathrm{Ker}(T)$ is its kernel and $\mathrm{Ran}(T)$ is its range. If $T : H_1 \to H_2$ is a bounded linear operator between F-Hilbert spaces H_i having inner products $\langle\cdot,\cdot\rangle_i$ $(i = 1, 2)$, its *adjoint* is the unique bounded linear operator $T^*\colon H_2 \to H_1$ for which

$$\langle T^*\psi, \phi\rangle_1 = \langle \psi, T\phi\rangle_2,$$

for all $\phi \in H_1, \psi \in H_2$. The bounded linear operator $U\colon H_1 \to H_2$ is said to be *unitary* if it is both an isometry and a co-isometry (i.e., U^* is also an isometry). Equivalently it is an isometric isomorphism for which $U^{-1} = U^*$.

If E is an F-Banach space, then $\mathcal{L}(E)$ will denote the algebra of all bounded linear operators on E. $\mathcal{L}(E)$ is a Banach space with respect to the operator norm

$$||T|| = \sup\{||Tx||; x \in H, ||x|| = 1\}.$$

Note that the algebra $M_n(F)$ may be realised as $\mathcal{L}(F^n)$. The (topological) dual space E' of E is the linear space of all bounded linear maps (often called linear functionals) from E to F. It is a Banach space with respect to the norm:

$$||l|| = \sup\{|l(x)|; x \in E, ||x|| = 1\}.$$

We typically use $\langle \cdot, \cdot \rangle$ to indicate the dual pairing between E' and E, so if $l \in E', x \in E$,

$$l(x) = \langle x, l \rangle.$$

If H is a Hilbert space and $x, y \in H$ are orthogonal so that $\langle x, y \rangle = 0$, we sometimes write $x \perp y$. In the main part of the book, we will always write $\langle \cdot, \cdot \rangle_{\mathbb{C}}$ as $\langle \cdot, \cdot \rangle$. It is perhaps worth emphasising that all inner products on complex vector spaces are linear on the left and conjugate-linear on the right, which is standard in mathematics (but not in physics). We use $\mathcal{H}$ instead of H to denote our Hilbert space whenever there is an operator called H (typically a quantum mechanical Hamiltonian) playing a role within that section of the text.

The inner product (scalar product) of $x, y \in \mathbb{R}^d$ is always written $x \cdot y$, and the associated norm is $|x| := \left(\sum_{i=1}^d x_i^2\right)^{\frac{1}{2}}$, for $x = (x_1, \ldots, x_d)$.

If $a, b \in \mathbb{R}$, then $a \wedge b := \min\{a, b\}$.

Throughout the book, we will use standard notation for partial differential operators. Let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be a *multi-index*, so that $\alpha \in (\mathbb{N} \cup \{0\})^d$. We define $|\alpha| = \alpha_1 + \cdots + \alpha_d$ and

$$D^\alpha = \frac{1}{i^{|\alpha|}} \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}}.$$

Similarly, if $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, then $x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$. For ease of notation, we will usually write ∂_i instead of $\frac{\partial}{\partial x_i}$ for $i = 1, \ldots, d$.

If S is a set, then we use ι for the identity mapping, $\iota(x) = x$, for all $x \in S$.

If f is a real or complex-valued function on $\mathbb{R}^d$ and $a \in \mathbb{R}^d$, $\tau_a f$ is the shifted function, defined by

$$(\tau_a f)(x) = f(x + a),$$

for all $x \in \mathbb{R}^d$.