1

Introduction

Analysis in metric spaces, in the sense that we are considering in this book, emerged as an independent research field in the late 1990s. Its origins lie in the search for an abstract context suitable to recover a substantial component of the classical Euclidean geometric function theory associated with quasiconformal and quasisymmetric mappings. Such a context, identified in the paper [125], consists of *doubling metric measure spaces* supporting a Poincaré inequality.

Over the past 15 years the subject of analysis in metric spaces has expanded dramatically. A significant part of that development has been a detailed study of abstract first-order Sobolev spaces and their relation to variational problems and partial differential equations as well as their role as a tool in, e.g., function theory, dynamics, and related fields. The subject advanced to the point where a careful treatment from first principles, in textbook form, appeared to be needed. This book is intended to serve that purpose.

The concept of an *upper gradient* plays a critical role in both the notion of Sobolev spaces considered in this book and the concomitant framework of metric measure spaces supporting a Poincaré inequality. This concept, also proposed originally in [125], provides an effective replacement for the gradient, or more precisely, the *norm of the gradient* of a smooth function. A nonnegative Borel function g (possibly taking the value $+\infty$) on a metric space (X, d) is said to be an upper gradient of a real-valued function u if the inequality

$$|u(x) - u(y)| \leq \int_\gamma g\, ds \tag{1.1}$$

is satisfied for all rectifiable curves γ joining x to y in X. Here the integral of g on the right-hand side of (1.1) is computed with respect to the arc length measure along γ induced by the metric d. We review the theory of path integrals along rectifiable curves in metric spaces in Chapter 5; Chapter 6 is devoted to the basic properties of upper gradients in metric spaces. It is worth emphasizing that no smoothness assumption on u is imposed a priori in the definition. (Indeed, it is not clear what such an assumption would entail in the metric space setting.) However, as we will see in this book, the existence of a well-behaved upper gradient for a function u necessarily implies certain regularity properties for u itself.

With the notion of upper gradient in hand it is natural to inquire about the existence of a theory of Sobolev spaces based on such gradients. The classical Sobolev space $W^{1,p}(\Omega)$, where Ω is a domain in $\mathbb{R}^n$, can be adapted to the setting of a metric measure space (X, d, μ) by introducing the space of p-integrable functions which admit a p-integrable upper gradient. The foundations for such a theory were laid in the thesis [247] and the accompanying paper [248]. In the literature this space is often referred to as the *Newtonian space* and denoted $N^{1,p}(X)$. This terminology highlights the essential role played by the upper gradient inequality (1.1), which in turn serves as an abstract counterpart of the fundamental theorem of calculus.

Chapters 7, 8, and 9 form the heart of this book. In these chapters we introduce and give a detailed study of the Sobolev space $N^{1,p}$. In these chapters, we show among other results that $N^{1,p}$ is a Banach function space, we study the pointwise properties of Sobolev functions (both scalar- and vector-valued), and we discuss the density of Lipschitz functions in the Sobolev space.

In this book we consistently employ the terminology *Sobolev space*, although we retain the notation $N^{1,p}(X)$ both in homage to the origins of the concept and to distinguish this space from other abstract versions of the classical Sobolev space. In Chapter 10 we review several alternate approaches to abstract Sobolev-type spaces on metric measure spaces. Under suitable assumptions, some of or all these spaces coincide, either as sets or (up to a linear isomorphism or even up to isometry) as Banach spaces.

One version of the classical Poincaré inequality on the Euclidean space $\mathbb{R}^n$ states that

$$\frac{1}{|B|}\int_B |u - u_B| \le Cr\,\frac{1}{|B|}\int_B |\nabla u|. \tag{1.2}$$

Here u denotes a C^∞-function on $\mathbb{R}^n$ and B denotes a ball of radius r. The notation $u_B = |B|^{-1}\int_B u$ denotes the mean value of u on B. The constant C depends only on the dimension n, i.e., it is independent of B and u.

Using the notion of the upper gradient one can reformulate the Poincaré inequality (1.2) in the metric measure space context, by replacing $|\nabla u|$ by any fixed upper gradient g of a given function u. Actually the story is more subtle. It follows trivially from Hölder's inequality that (1.2) implies the corresponding inequality where the integral on the right-hand side is replaced by the L^p-norm of $|\nabla u|$ with respect to the Lebesgue measure on B (normalized by the volume of B as in (1.2)). Moreover, one can replace the ball B by any larger concentric ball λB ($\lambda > 1$), at the cost of possibly changing the constant C. We say that a metric measure space (X, d, μ) *supports a weak* p*-Poincaré inequality* if there exist constants $C > 0$ and $\lambda \geq 1$ such that the inequality

$$\frac{1}{\mu(B)} \int_B |u - u_B| \, d\mu \leq Cr \left(\frac{1}{\mu(\lambda B)} \int_{\lambda B} g^p \, d\mu \right)^{1/p} \tag{1.3}$$

holds for all balls B in X and all function–upper-gradient pairs (u, g). As before, r denotes the radius of B while λB denotes the ball with the same center as B and with radius λr.

The importance of the abstract Poincaré inequality (1.3) lies in the fact that it imposes an additional relation between functions and their upper gradients, at the level of the volume measure μ rather than at the level of the length measure along curves. The length–volume principle (usually known as the *length–area principle*) lies at the core of classical Euclidean geometric function theory. In our setting the interplay between the upper gradient inequality (1.1) and the Poincaré inequality (1.3) is a principal driving force. When coupled with the *doubling condition* for the measure μ (namely, the assumption that $\mu(2B) \leq C\mu(B)$ for all balls B in X, where the constant C is independent of B), the Poincaré inequality becomes a powerful tool with both analytic and geometric consequences.

The reader may wonder why we complicate the story by distinguishing the Poincaré inequality according to the value of the exponent p as well as by allowing for dilated balls λB in the definition. In Euclidean space, as already observed, the Poincaré inequality holds with $p = 1$ and $\lambda = 1$ (and this is the strongest form of the inequality). In an abstract setting, it is not necessarily the case that a space supporting a Poincaré inequality for some $1 \leq p < \infty$ and with some dilation constant $\lambda \geq 1$ necessarily supports a Poincaré inequality for better choices of this data. Under rather mild conditions, however, the dilation parameter λ can always be chosen to be 1. We discuss this and other self-improvement phenomena relating to Sobolev–Poincaré inequalities in Chapter 9.

It is a much deeper fact of the theory that if the underlying metric space is complete and the measure μ is doubling then the exponent p on the right-hand side of (1.3) can be improved. In other words, if such a space (X, d, μ) supports a p-Poincaré inequality for some $p > 1$ then it supports a q-Poincaré inequality for some $1 \leq q < p$. This fact, due to Keith and Zhong, is a highlight of the modern theory of analysis on metric spaces. Chapter 12 contains a detailed and self-contained proof of the Keith–Zhong theorem as well as a discussion of its numerous implications and corollaries. Examples of doubling spaces supporting a p-Poincaré inequality for some, but not all, values of p in the range $[1, \infty)$ are described in Chapters 13 and 14.

Of comparable importance is the landmark theorem of Cheeger on the almost everywhere differentiability of Lipschitz functions on doubling spaces supporting a Poincaré inequality. This result, an abstract reformulation of the famous Rademacher differentiation theorem for Euclidean Lipschitz functions, demonstrates that doubling metric measure spaces supporting a Poincaré inequality possess a rich infinitesimal "linear" structure not immediately apparent from the indefinition. Indeed, on such spaces it is possible to define not only the norm of the gradient of a Lipschitz function but (in a suitable sense) the *gradient* or differential itself, acting as a linear operator. Chapter 13 contains a proof of Cheeger's differentiation theorem.

One of our aims in preparing this book has been to present self-contained proofs of these two key theorems of Keith and Zhong and Cheeger.

Another major theme of this book is our consistent emphasis on the class of *vector-valued functions*, that is to say, functions taking values in a Banach space V. The integrability theory for vector-valued functions goes back to the work of Bochner and Pettis; we review this theory in Chapter 3. Our standard setting is that of V-valued Bochner integrable functions u defined on a metric measure space (X, d, μ). (Note, however, that the upper gradients of such functions u, as analogs of the norm of the classical gradient, remain real-valued functions.) The theory of first-order Sobolev spaces is, with a few notable expections, no more difficult to develop in the vector-valued case than in its scalar-valued counterpart. Moreover, there are important reasons why one wishes to have a theory in such a context. Every metric space admits an isometric embedding into some Banach space. (See Chapter 4 for a summary of classical embedding and extension theorems.) Taking advantage of such embeddings one can define metric space-valued Sobolev mappings. The space of Sobolev mappings from a metric measure space (X, d, μ) into another metric space (Y, d') plays a key analytic role in the theory of quasisymmetric maps as well as in nonlinear geometric variational problems. While we do not investigate those subjects in this book, we can remark that the analytic definition of quasisymmetric maps in terms of metric space-valued Sobolev mappings, as developed in our paper [129], was a primary impetus for this book. A brief survey of the theory of quasiconformal and quasisymmetric mappings on metric spaces can be found in Section 14.1.

In Chapter 14 we describe various examples of metric measure spaces supporting a Poincaré inequality and, although we do not provide proofs of the relevant inequality for these examples, we do give copious references to the literature in case the reader wishes to pursue such matters further. It is also useful to know that the collection of doubling metric measure spaces supporting a Poincaré inequality, with uniform constants, is closed under a suitable notion of convergence (e.g., convergence in the *Gromov–Hausdorff* sense). We discuss Gromov–Hausdorff convergence and prove the preceding claim in Chapter 11. This observation expands the class of example spaces for our theory by including suitable Gromov–Hausdorff limit spaces.

The following references are recommended to readers who wish to learn more about the subject. The short books [120] and [8] are good introductions to the field of analysis in metric spaces. Hajłasz's survey articles [109] and [112] focus specifically on the theory of Sobolev spaces on metric spaces; these two articles are well suited for readers wishing to learn more about alternative notions of Sobolev spaces as discussed in Chapter 10. For a general historical survey of nonsmooth calculus, see [122]. The book by A. and J. Björn [31] is a comprehensive treatment of nonlinear potential theory, especially the theory of *p-harmonic functions* on metric measure spaces; this book serves as a valuable counterpart to the present volume. Other topics closely related to the subject matter of this book, and that are currently under active study, include abstract notions of curvature (as in the books [8] and [276]) and analysis on fractals (as in the book [155]).

This book is intended as a graduate textbook. We have endeavored to include detailed proofs of virtually all the major results and to present the material in such a way as to minimize the assumed background. However, prior knowledge of abstract measure theory and functional analysis, at the level of a standard introductory graduate course, is highly recommended. We review the basic tools of functional analysis needed for this book in Chapter 2, while in Chapter 3 we review the foundations of Borel and Radon measures, the integration of Banach space-valued functions, and the basic tools of harmonic analysis such as the Hardy–Littlewood maximal function. A prior exposure to Sobolev spaces (e.g., as can be found in a graduate partial differential equation course) could help the reader to place the topics of this book in a broader context.

Throughout this book, we let C denote any positive constant whose particular value is not of interest to us; thus, even within the same line, two occurrences of C may refer to two different values. However, C will always be assumed to be a positive constant.

Our style of exposition has undoubtedly been influenced by the works of our mathematical fathers and grandfathers, including Olli Martio, Olli

Lehto, and Rolf Nevanlinna. Besides this we wish to acknowledge Jussi Väisälä, whose lecture notes on quasiconformal mappings attracted each of us to the subject. Finally, we have benefited tremendously from the inspiring atmosphere generated by Lois and Fred Gehring, and from the mentoring which we have all received from Fred. We dedicate this book with great appreciation to his memory.

2

Review of basic functional analysis

The theory of Sobolev spaces as developed in this book requires only a small amount of elementary functional analysis. In this chapter we present the required background material. For the sake of completeness, we have included proofs for all but the most standard facts. Anyone with a good working knowledge of analysis can safely skip this chapter. Alternatively, one can quickly glance through the chapter to check the notation and return to it later as needed.

We assume that the reader is familiar with the basic measure theory of real-valued functions and Lebesgue integration. The integration theory for Banach space-valued functions will be developed in Chapter 3.

2.1 Normed and seminormed spaces

Let V be a vector space over the real numbers. A *norm* on V is a function

$$|\cdot| : V \to \mathbb{R}$$

that satisfies the following three conditions:

$$|v| > 0 \qquad \text{for all } v \in V \setminus \{0\}, \tag{2.1.1}$$

$$|\lambda v| = |\lambda|\,|v| \qquad \text{for all } v \in V \text{ and } \lambda \in \mathbb{R}, \tag{2.1.2}$$

$$|v + w| \leq |v| + |w| \qquad \text{for all } v, w \in V. \tag{2.1.3}$$

Here and throughout this book, $|\lambda|$ denotes the absolute value of a real number λ. The notational similarity between absolute values and general norms should not cause any confusion.

If $|\cdot|$ is a norm and $v \in V$, it follows from the definition that $|v| \geq 0$ and that $|v| = 0$ if and only if $v = 0$. A function $|\cdot| : V \to \mathbb{R}$ is called a *seminorm* on V if it satisfies (2.1.2), (2.1.3), and in place of (2.1.1) the following weaker version:

$$|v| \geq 0 \qquad \text{for all } v \in V. \tag{2.1.4}$$

If $|\cdot|$ is a norm on V, the pair $(V, |\cdot|)$ is called a *normed space*. Analogously, $(V, |\cdot|)$ is a *seminormed space* if $|\cdot|$ is a seminorm.

The n-dimensional space $\mathbb{R}^n$, $n \geq 1$, is most commonly equipped with the *Euclidean norm*

$$|x| = (x_1^2 + \cdots + x_n^2)^{1/2}, \quad x = (x_1, \ldots, x_n). \tag{2.1.5}$$

We always assume, unless otherwise explicitly stated, that $\mathbb{R}^n$ comes equipped with the norm (2.1.5). There are, however, many other norms in $\mathbb{R}^n$. For $1 \leq p \leq \infty$ we have the p-norms

$$|x|_p := (|x_1|^p + \cdots + |x_n|^p)^{1/p}, \qquad 1 \leq p < \infty, \tag{2.1.6}$$

and

$$|x|_\infty := \max\{|x_1|, \ldots, |x_n|\}. \tag{2.1.7}$$

Thus $|x| = |x|_2$ for $x \in \mathbb{R}^n$.

The norms $|\cdot|_p$ can be defined as extended real-valued functions on the vector space of infinite sequences $\mathbb{R}^\infty := \{(x_1, x_2, \ldots) : x_i \in \mathbb{R}\}$ in an obvious way,

$$|x|_p := \left(\sum_{i=1}^\infty |x_i|^p\right)^{1/p}, \qquad |x|_\infty := \sup\{|x_i| : i = 1, 2, \ldots\}. \tag{2.1.8}$$

Then a family of norms can be defined by restricting $|x|_p$ to the vector subspace of $\mathbb{R}^\infty$ consisting of those $x \in \mathbb{R}^\infty$ for which $|x|_p < \infty$. In this way we construct the l^p*-spaces*,

$$l^p = l^p(\mathbb{N}) := \{x \in \mathbb{R}^\infty : |x|_p < \infty\}, \qquad 1 \leq p \leq \infty. \tag{2.1.9}$$

More generally, let (X, μ) be a measure space, where μ is a measure on the set X (see Section 3.1 for a review of the basic terminology) and define, for $1 \leq p < \infty$ and measurable $f : X \to [-\infty, \infty]$,

$$\|f\|_p := \left(\int_X |f|^p \, d\mu\right)^{1/p}. \tag{2.1.10}$$

Then $\|\cdot\|_p$ is a seminorm on the vector space of measurable functions f for which $\|f\|_p < \infty$. It is not always a norm, for the integral in (2.1.10) vanishes whenever f vanishes almost everywhere. If we identify two functions that agree almost everywhere then, for the resulting equivalence classes $[f]$, we can define $\|[f]\|_p$ unambiguously via (2.1.10) by using a representative. In this way we arrive at the L^p*-spaces*,

$$L^p = L^p(X) = L^p(X, \mu), \qquad 1 \le p < \infty, \tag{2.1.11}$$

consisting of the equivalence classes $[f]$ of measurable functions on X with $\|[f]\|_p < \infty$. It is customary in L^p-theory to speak about *functions* in L^p rather than equivalence classes, and to use the notation f and $\|f\|_p$ rather than $[f]$ and $\|[f]\|_p$. We will follow the same practice. In the theory of Sobolev spaces, the issue of the identification of functions arises in a more subtle way; this will be discussed in detail in Chapters 5 and 6. Functions in $L^p(X)$ are also referred to as p*-integrable functions* on X.

The *sup norm* for a measurable function $f : X \to [-\infty, \infty]$ is given by

$$\|f\|_\infty := \sup\{\lambda \in \mathbb{R} : \mu(\{x \in X : |f(x)| > \lambda\}) \neq 0\}. \tag{2.1.12}$$

Upon following the preceding identification convention for functions, we obtain the normed space

$$L^\infty = L^\infty(X) = L^\infty(X, \mu). \tag{2.1.13}$$

This is the space of *essentially bounded functions*, consisting of those (equivalence classes of) measurable functions for which the expression $\|f\|_\infty$ is finite.

In the case when $X = \mathbb{N}$ and μ is the counting measure, we recover the l^p-spaces as in (2.1.9).

For an arbitrary set A, with no assigned measure, one can define a normed space

$$l^\infty(A) \tag{2.1.14}$$

consisting of all bounded functions $f : A \to \mathbb{R}$ with the norm

$$\|f\|_\infty := \sup_{a \in A} |f(a)|. \tag{2.1.15}$$

We will use the shorthand notation $l^\infty = l^\infty(\mathbb{N})$, $\|x\|_\infty = |x|_\infty$ for $x \in \mathbb{R}^\infty$, which is in accordance with (2.1.8) and (2.1.9).

Remark 2.1.16 The procedure of passing to the equivalence classes of functions in L^p-spaces is an example of a general procedure whereby a seminormed space can be turned into a normed space. To wit, let $(V, |\cdot|_S)$ be a seminormed space. For $v \in V$ we consider the *equivalence class* $[v]$ given by the equivalence relation $\sim$, where $v \sim w$ if and only if $|v - w|_S = 0$. By setting

$$\big|[v]\big| = |v|_S \tag{2.1.17}$$

we obtain a norm in the vector space of equivalence classes $[v]$. Put differently, if V_S denotes the vector subspace of V consisting of those vectors v for which $|v| = 0$ then the map $|\cdot|_S : V \to \mathbb{R}$ factors through the canonical projection $V \to V/V_S$ as a norm $|\cdot| : V/V_S \to \mathbb{R}$.

Lebesgue measure We denote the n-dimensional Lebesgue measure on $\mathbb{R}^n$ by m_n and the corresponding Lebesgue spaces by $L^p(\mathbb{R}^n)$. More generally, if $A \subset \mathbb{R}^n$ is a Lebesgue measurable set then the short notation

$$L^p(A) = L^p(A, m_n)$$

is used, where m_n is restricted to A in a natural manner.

Metric spaces

A *metric space* is a pair (X, d), where X is a set and $d : X \times X \to [0, \infty)$ is a function called a *distance* or *metric* satisfying the following three conditions:

$$d(x, y) = d(y, x) \qquad \text{for all } x, y \in X, \tag{2.1.18}$$

$$d(x, y) = 0 \qquad \text{if and only if } x = y, \tag{2.1.19}$$

$$d(x, y) \le d(x, z) + d(z, y) \qquad \text{for all } x, y, z \in X. \tag{2.1.20}$$

Both (2.1.3) and (2.1.20) are commonly referred to as the *triangle inequality*. We assume that the reader is familiar with the basic theory of metric spaces, including standard topological notions such as completeness and compactness. A reasonable discussion of this basic theory can be found in [214].

It follows from the definitions that every normed space $(V, |\cdot|)$ is naturally a metric space with distance function $d(v, w) = |v - w|$. Unless otherwise stated, all topological notions on a normed space $V = (V, |\cdot|)$ will be based on this metric. For example, "the sequence (v_i) converges to v in V", or "$v_i \to v$ in V", means $\lim_{i\to\infty} |v_i - v| = 0$. We will, however, consider other modes of convergence in V later (see Section 2.3).

A metric space is *separable* if it possesses a countable dense subset. A normed space is said to be separable if it is separable as a metric space.

The space $L^p(X)$ for $1 \le p < \infty$ is separable under some mild conditions on the measure space $X = (X, \mu)$. For example, $L^p(\mathbb{R}^n)$ is separable for $1 \le p < \infty$. (See Proposition 3.3.49 for a statement in the main context of this book.) On the other hand, the space $L^\infty(X)$ is rarely separable, and $l^\infty(A)$ is separable if and only if A is a finite set.

Banach spaces

A normed space $(V, |\cdot|)$ is said to be a *Banach space* if it is complete as a metric space. We also use the self-explanatory term *complete norm* in this case.