FRACTALS AND SELF SIMILARITY

JOHN E. HUTCHINSON

This is a retyped (TeX’d) version of the article from Indiana University Mathematics Journal 30 (1981), 713–747 (with some minor formatting changes, a few old “typos” corrected, and hopefully few new ones introduced).

The original preprint appeared as Research Report No. 31-1979, Department of Pure Mathematics, Faculty of Science, Australian National University.

Contents

1. Introduction 2
2. Preliminaries 3
2.1. Sequences of Integers 3
2.2. Maps in Metric Spaces 4
2.3. Similitudes 4
2.4. Hausdorff Metric 6
2.5. Measures 6
2.6. Hausdorff Measure 7
2.7. Geometric Measure Theory 8
3. Invariant Sets 10
3.1. Elementary Proof of Existence and Uniqueness, and Discussion of Properties 10
3.2. Convergence in the Hausdorff Metric 12
3.3. Examples 13
3.4. Remark 14
3.5. Parametrised Curves 14
4. Invariant Measures 15
4.1. Motivation 15
4.2. Definitions 16
4.3. The L metric 16
4.4. Existence and Uniqueness 16
4.5. Different Sets of Similitudes Generating the Same Set 17
5. Similitudes 18
5.1. Self-Similar Sets 18
5.2. Open Set Condition 18
5.3. Existence of Self Similar Sets. 19
5.4. Purely Unrectifiable Sets. 21
5.5. Parameter Space 24
6. Integral Flat Chains 24
6.1. The $F$-metric. 24
6.2. The $C$-metric 25
6.3. Invariant Chains 26
References 27
1. Introduction

Sets with non-integral Hausdorff dimension (2.6) are called fractals by Mandelbrot. Such sets, when they have the additional property of being in some sense either strictly or statistically self-similar, have been used extensively by Mandelbrot and others to model various physical phenomena (c.f. [MB] and the references there). However, these notions have not so far been studied in a general framework.

In this paper we set up a theory of (strictly) self-similar objects, in a subsequent paper we analyse statistical self-similarity.

We now proceed to indicate the main results. The reader should refer to the examples in 3.3 for motivation. We say the compact set $K \subset \mathbb{R}^n$ is invariant if there exists a finite set $\mathcal{S} = \{S_1, \ldots, S_N\}$ of contraction maps on $K \subset \mathbb{R}^n$ such that

$$K = \bigcup_{i=1}^{N} S_i K.$$

In such a case we say $K$ is invariant with respect to $\mathcal{S}$. Often, but not always, the $S_i$ will be similitudes, i.e. a composition of an isometry and a homothety (2.3).

In [MB], and in the case the $S_i$ are similitudes, such sets are constructed by an iterative procedure using an “initial” and a “standard” polygon. However, here we need to consider instead the set $\mathcal{S}$.

It turns out, somewhat surprisingly at first, that the invariant set $K$ is determined by $\mathcal{S}$. In fact, for given $\mathcal{S}$ there exists a unique compact set $K$ invariant with respect to $\mathcal{S}$. Furthermore, $K$ is the limit of various approximating sequences of sets which can be constructed from $\mathcal{S}$.

More precisely we have the following result from 3.1(3), 3.2.

(1) Let $X = (X,d)$ be a complete metric space and $\mathcal{S} = \{S_1, \ldots, S_N\}$ be a finite set of contraction maps (2.2) on $X$. Then there exists a unique closed bounded set $K$ such that $K = \bigcup_{i=1}^{N} S_i K$. Furthermore, $K$ is compact and is the closure of the set of fixed points $s_{i_1 \ldots i_p}$ of finite compositions $S_{i_1} \circ \cdots \circ S_{i_p}$ of members of $\mathcal{S}$.

For arbitrary $A \subset X$ let $\mathcal{S}(A) = \bigcup_{i=1}^{N} S_i A$, $\mathcal{S}^p(A) = \mathcal{S}(\mathcal{S}^{p-1}(A))$. Then for closed bounded $A$, $\mathcal{S}^p(A) \to K$ in the Hausdorff metric (2.4).

The compact set $K$ in (1) is denoted $|\mathcal{S}|$. $|\mathcal{S}|$ supports various measures in a natural way. We have the following from 4.4.

(2) In addition to the hypotheses of (1), suppose $\rho_1, \ldots, \rho_N \in (0,1)$ and $\sum_{i=1}^{N} \rho_i = 1$. Then there exists a unique Borel regular measure $\mu$ of total mass 1 such that $\mu = \sum_{i=1}^{N} \rho_i S_i \#(\mu)$. Furthermore $\text{sp}t(\mu) = |\mathcal{S}|$.

The measure $\mu$ is denoted $||\mathcal{S}|, \rho||$.

The set $|\mathcal{S}|$ will not normally have integral Hausdorff dimension. However, in case $(X,d)$ is $\mathbb{R}^n$ with the Euclidean metric, $|\mathcal{S}|$ can often be treated as an $m$-dimensional object, $m$ an integer, in the sense that there is a notion of integration of $C^\infty$ $m$-forms over $|\mathcal{S}|$. In the language of geometric measure theory (2.7), $|\mathcal{S}|$ supports an $m$-dimensional integral flat chain. The main result here is 6.3(3).

Now suppose $(X,d)$ is $\mathbb{R}^n$ with the Euclidean metric, and the $S_i \in \mathcal{S}$ are similitudes. Let $\text{Lip} S_i = r_i$ (2.2) and let $D$ be the unique positive number for which $\sum_{i=1}^{N} r_i^D = 1$. Then $D$ is called the similarity dimension of $\mathcal{S}$, a term coined by Mandelbrot. In case a certain “separation” condition holds, namely the open set condition of 5.2(1), one has the following consequences from 5.3(1) (see 2.6(1), (3) for notation).

(3) \begin{enumerate}
  \item $D =$Hausdorff dimension of $|\mathcal{S}|$ and $0 < \mathcal{H}^D(|\mathcal{S}|) < \infty$,
  \item $\mathcal{H}^D(S_i|\mathcal{S}| \cap S_j|\mathcal{S}|) = 0$ if $i \neq j$.
\end{enumerate}
(iii) there exist $\lambda_1, \lambda_2$ such that for all $k \in |S|$, 
$$0 < \lambda_1 \leq \theta^D([S], k) \leq \theta^{\ast D}(|S|, k) \leq \lambda_2 < \infty,$$
(iv) $\|S, \rho\| = [\mathcal{H}^D([S])]^{-1} \mathcal{H}^D[|S|]$ if $\rho = r^D$.

A result equivalent to (3)(i) was first proved by Moran in [MP].

With a stronger separation condition we prove in 5.4(1) that for suitable $m, |S|$ meets no $m$-dimensional $C^1$ manifold in a set of strictly positive $\mathcal{H}^m$ measure. In the notation of [FH], $|S|$ is purely $(\mathcal{H}^m, m)$ unrectifiable.

In the case of similitudes in $\mathbb{R}^n$, it is possible to parametrise invariant sets by points in a $C^\infty$ manifold (5.5).

My special thanks to F. J. Almgren, Jr., for making possible my stay at Princeton University, for suggesting I begin this study, and for his continuing comments and enthusiasm. I would also like to thank L. Simon for advice and for his invitation to Melbourne University. Helpful suggestions came from members of the Princeton and Melbourne seminars, especially R. Hardt, V. Scheffer and B. White.

Finally I wish to thank Benoit Mandelbrot, from whose ideas this paper developed.

2. Preliminaries

$(X, d)$ is always a complete metric space, often Euclidean space $\mathbb{R}^n$ with the Euclidean metric.

$$B(a, r) = \{x \in X : d(a, x) \leq r\},$$
$$U(a, r) = \{x \in X : d(a, x) < r\}.$$

If $A \subset X$, then $\overline{A}$ is the closure of $A$, $A^\circ$ is the interior, $\partial A$ is the boundary, and $A^c$ is the complement $X \sim A$.

A $C^1$ function is one whose first partial derivatives exist and are continuous. A $C^\infty$ function is a function having partial derivatives of all orders.

A $C^1$ manifold in $\mathbb{R}^n$ will mean a continuously differentiable embedded submanifold having the induced topology from $\mathbb{R}^n$.

A proper function is a function for which the inverse of every compact set is compact.

2.1. Sequences of Integers. $P = \{1, 2, \ldots\}$ is the set of positive integers. $N \in P$, $N \geq 2$ is usually fixed.

(1) Ordered $p$-tuples are denoted $(i_1, \ldots, i_p)$, where usually each $i_j \in \{1, \ldots, N\}$. We write $\alpha \prec \beta$ if $\alpha$, $\beta$ are $p$-tuples with $\alpha$ an initial segment of $\beta$, i.e. $\alpha = (i_1, \ldots, i_p)$ and $\beta = (i_1, \ldots, i_p, i_{p+1}, \ldots, i_{p+q})$ for some $q \geq 0$. $\alpha \preceq \beta$ means $\alpha \prec \beta$ and $\alpha \neq \beta$.

(2) $C(N)$, the Cantor set on $N$ symbols, is the set of maps (i.e. sequences) $\alpha : P \to \{1, \ldots, N\}$. Thus $C(N) = \prod_{p=1}^{\infty} \{1, \ldots, N\}$. We write $\alpha_p$ for $\alpha(p)$. A typical element of $C(N)$ is often written $\alpha_1 \ldots \alpha_p$, $i_1 \ldots i_p \ldots$. We extend the notation $\alpha \prec \beta$ to the case $\alpha = (i_1, \ldots, i_p)$ and $\beta = (i_1, \ldots, i_p, i_{p+1}, \ldots, i_{p+q}) \in C(N)$.

(3) If $i \in \{1, \ldots, N\}$ and $\alpha = (i_1, \ldots, i_p)$ is a $p$-tuple, then $i\alpha = (i, i_1, \ldots, i_p)$ is just concatenation of $i$ and $\alpha$. Similarly if $\alpha \in C(N)$ then $i\alpha = i\alpha_1 \ldots \alpha_{p-1}\alpha_p \ldots$. Likewise if $\beta$ is a $q$-tuple and $\alpha$ is a $p$-tuple or $\alpha \in C(N)$ we form $\beta\alpha$ in the obvious way.

The $i$th shift operator $\sigma_i = C(N) \to C(N)$ is given by $\sigma_i(\alpha) = i\alpha$.

(4) $C(N)$ is given the product topology (also called the weak topology) induced from the discrete topology on each factor $\{1, \ldots, N\}$. Thus a sub-basis of open sets is given by sets of the form $\{\alpha : \alpha_p = i\}$ where $p \in P$, $i \in \{1, \ldots, N\}$. $C(N)$ is compact.
(5) By $\hat{1}_1\ldots\hat{1}_p$ we mean the infinite sequence $i_1\ldots i_p i_1\ldots i_p \ldots i_1 \ldots i_p \ldots \in C(N)$. Thus $i_1\ldots i_{p+q}^+\ldots\hat{1}_{p+q}$ may be regarded as the general rational element of $C(N)$.

(6) The set $I$ will always be a finite set of finite ordered tuples (of not necessarily equal length) from $\{1, \ldots, N\}$.

$I = \{\alpha_1\ldots\alpha_q : \alpha_i \in I\} \subset C(N)$, where we are concatenating finite ordered tuples in the obvious way. Thus if $I = \{1, \ldots, N\}$ then $I = C(N)$. If $I = \{(1,2), (1)\}$ then $2\alpha_2 \ldots \alpha_p \notin I$, etc.

(7) For $\alpha$ an ordered tuple, let $\alpha^* = \{\beta \in C(N) : \alpha \prec \beta\}$.

We say $I$ is secure if for every $\beta \in C(N)$ there exists $\alpha \in I$ such that $\alpha \prec \beta$. This is equivalent to: for every $p$-tuple $\beta$ with $p = \max\{\text{length} \cdot \alpha : \alpha \in I\}$, there exists $\alpha \in I$ such that $\alpha \prec \beta$. Since $I$ is finite there is an obvious algorithm to check if $I$ is secure.

We say $I$ is tight if for every $\beta \in C(N)$ there exists exactly one $\alpha \in I$ such that $\alpha \prec \beta$. Again one can always check, in a finite number of steps, if $I$ is tight.

(8) Proposition

(i) The following are equivalent

(1) $I = C(N)$,

(2) $C(N) = \bigcup_{\alpha \in I} \alpha^*$,

(3) $I$ is secure.

(ii) The following are equivalent:

(1) Each member of $C(N)$ has a unique decomposition of the form $\alpha_1\ldots\alpha_q\ldots$, with $\alpha_i \in I$,

(2) $C(N) = \bigvee_{\alpha \in I} \alpha^*$ (disjoint union),

(3) $I$ is tight.

Proof. In both cases the implications (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (i) are clear. \hfill $\square$

(9) One can check that $I$ is tight iff $I$ is essential and satisfies the tree condition, in the sense of [OP, III].

2.2. Maps in Metric Spaces. If $F : X \to X$, then we define the Lipschitz constant of $F$ by

$$\text{Lip} F = \sup_{x \neq y} \frac{d(F(x), F(y))}{d(x, y)}.$$  

Of course if $\text{Lip} F = \lambda$, then $d(F(x), F(y)) \leq \lambda d(x, y)$ for all $x, y \in X$, and moreover $\text{Lip} F$ is the least such $\lambda$. We say $F$ is Lipschitz if $\text{Lip} F < \infty$ and $F$ is a contraction if $\text{Lip} F < 1$.

(1) It is a standard fact that every contraction map (in a complete metric space) has a unique fixed point.

(2) Definition. Suppose $S = \{S_1, \ldots, S_N\}$ is a finite family of maps $S_i : X \to X$. Then $S_{1\ldots N} = S_1 \circ \cdots \circ S_N$.

2.3. Similitudes. $S : X \to X$ is a similitude if $d(S(x), S(y)) = rd(x, y)$ for all $x, y \in X$ and some fixed $r$.

- $\mu_r : \mathbb{R}^n \to \mathbb{R}^n$ is the homothety $\mu_r(x) = rx \ (r \geq 0)$.
- $\tau_b : \mathbb{R}^n \to \mathbb{R}^n$ is the translation $\tau_b(x) = x - b$.

(1) Proposition. $S : \mathbb{R}^n \to \mathbb{R}^n$ is a similitude iff $S = \mu_r \circ \tau_b \circ O$ for some homothety $\mu_r$, translation $\tau_b$, and orthonormal transformation $O$. 

Proof. The “only if” is clear.

Conversely, let $S$ be a similitude, $\text{Lip } S = r \neq 0$. Let $g(x) = r^{-1}(S(x) - S(0))$. Then $g$ is an isometry fixing 0.

Since
\[
(x, y) = \frac{1}{2}[||x||^2 + ||y||^2 - ||x - y||^2]
\]
\[
= \frac{1}{2}[d(0, x)^2 + d(0, y)^2 - d(x, y)^2],
\]
it follows $g$ preserves inner products.

Let $\{e_i : 1 \leq i \leq N\}$ be an orthonormal basis for $\mathbb{R}^n$. Then $\{g(e_i) : 1 \leq i \leq N\}$ is also on orthonormal basis, and hence
\[
g(x) = \sum_{i=1}^{n} (g(x), g(e_i)) g(e_i) = \sum_{i=1}^{n} (x, e_i) g(e_i),
\]
since $g$ preserves inner products. It follows $g$ is linear and so is an orthonormal transformation.

Since
\[
S(x) = r g(x) + S(0) = r(g(x) + r^{-1}S(0)),
\]
it follows
\[
S = \mu_r \circ \tau_{-r^{-1}S(0)} \circ g,
\]
and we are done.

(2) **Remark.** The same proof works in a Hilbert space to show that $S$ is a similitude iff $S = \mu_r \circ \tau_{b} \circ O$, where now $O$ is a unitary transformation.

(3) **Convention.** For the rest of this paper, unless mentioned otherwise, all similitudes are contractions.

(4) Returning to the case $(\mathbb{R}^n, d)$, let the similitude $S$ have fixed point $a$, let $\text{Lip } S = r$, and let $O$ be the orthonormal transformation given by $O(x) = r^{-1}[S(x + a) - a]$ (orthonormal since the origin is clearly fixed and $O$ is clearly an isometry, now use (1)).

Then
\[
S(x + a) = rO(x) + a,
\]
so
\[
S(x) = rO(x - a) + a,
\]
and hence
\[
S = \tau_a^{-1} \circ \mu_r \circ O \circ \tau_a = (\tau_a^{-1} \circ \mu_r \circ \tau_a) \circ (\tau_a^{-1} \circ O \circ \tau_a),
\]
so that $S$ may be conveniently thought of as an orthonormal transformation about $a$ followed by a homothety about $a$. We write
\[
S = (a, r, O)
\]
and say that $S$ is in **canonical form.** $a$ and $r$ are uniquely determined by $S$, and so is $O$ if $r \neq 0$.

If $S_1 = (a_1, r_1, O_1)$, $S_2 = (a_2, r_2, O_2)$, then $S_1 \circ S_2 = (a, r, O)$ where $r = r_1 r_2$ and $O = O_1 \circ O_2$. However the expression for $a$ is not as simple, a calculation gives
\[
a = a_2 + (I - r_1 r_2 O_1 O_2)^{-1}(I - r_1 O_1)(a_2 - a_1).
2.4. Hausdorff Metric. If \( x \in X \), \( A \subset X \), define the distance between \( x \) and \( A \) by
\[
d(x, A) = \inf \{d(x, a) : a \in A\}.
\]
If \( A \subset X \), \( \varepsilon > 0 \), define the \( \varepsilon \)-neighbourhood of \( A \) by
\[
A_\varepsilon = \{x \in X : d(x, A) < \varepsilon\}.
\]
Thus \( A \subset A_\varepsilon \).

Let \( B \) be the class of non-empty closed bounded subsets of \( X \). Let \( C \) be the class of non-empty compact subsets.

Define the Hausdorff metric \( \delta \) on \( B \) by
\[
\delta(A, B) = \sup \{d(a, B), d(b, A) : a \in A, b \in B\}.
\]
Thus \( \delta(A, B) < \varepsilon \) iff \( A \subset B_\varepsilon \) and \( B \subset A_\varepsilon \). It is easy to check that \( \delta \) is a metric on \( B \).

It follows from [FH, 2.10.21] that \( (B, \delta) \) is a complete metric space. It also follows that if \( K \subset X \) is compact, then \( C \cap \{A : A \subset K\} \) is compact.

Some elementary properties of \( \delta \) which we will use are: let \( F : X \to X \), then
\[
(i) \quad \delta(F(A), F(B)) \leq \operatorname{Lip}(F) \delta(A, B),
(ii) \quad \delta(\bigcup_{i \in I} A_i, \bigcup_{i \in I} B_i) \leq \sup_{i \in I} \delta(A_i, B_i).
\]

2.5. Measures.

(1) A measure \( \mu \) on a set \( X \) is a map \( \mu : \mathcal{P}(X) = \{A : A \subset X\} \to [0, \infty] \) such that
\[
(i) \quad \mu(\emptyset) = 0,
(ii) \quad \mu\left(\bigcup_{i=1}^\infty E_i\right) \leq \sum_{i=1}^\infty \mu(E_i), \quad E_i \subset X.
\]
It follows \( A \subset B \) implies \( \mu(A) \leq \mu(B) \). Thus \( \mu \) is what is often called an outer measure. One says \( A \) is measurable iff \( \mu(T) = \mu(T \cap A) + \mu(T \sim A) \) for all \( T \subset X \).

The family of measurable sets forms a \( \sigma \)-algebra, \( \mu \) is a finite measure if \( \mu(X) < \infty \).

If \( A \subset X \), \( \mu[A] \) is the measure defined by \( \mu[A;E] = \mu(A \cap E) \).

For now on, \( X = (X, d) \) is a complete metric space. One says that \( \mu \) is Borel regular iff all Borel sets are measurable and for each \( A \subset X \) there exists a Borel set \( B \supset A \) with \( \mu(A) = \mu(B) \). If \( \mu \) is finite and Borel regular, it follows from [FH, 2.2.2] that for arbitrary Borel sets \( E \subset X \),
\[
(i) \quad \mu(E) = \sup \{\mu(K) : E \supset K \text{ closed}\},
(ii) \quad \mu(E) = \inf \{\mu(V) : E \subset V \text{ open}\}.
\]

(2) We define the support of \( \mu \) to be the closed set
\[
spt \mu = X \sim \bigcup \{V : V \text{ open, } \mu(V) = 0\}.
\]
Define the mass of \( \mu \) by
\[
M(\mu) = \mu(X).
\]

Define \( \mathcal{M} \) to be the set of Borel regular measures having bounded support and finite mass.

Define \( \mathcal{M}^1 = \{\mu \in \mathcal{M} : M(\mu) = 1\} \).

For \( a \in X \) define \( \delta_a = [a][a] \in \mathcal{M}^1 \) by \( \delta_a(A) = 1 \) if \( a \in A \), \( \delta_a(A) = 0 \) if \( a \notin A \).

(3) Let \( BC(X) = \{f : X \to R : f \text{ is continuous and bounded on bounded subsets}\} \).

For \( \mu \in \mathcal{M} \), \( \phi \in BC(X) \), define \( \mu(\phi) = \int \phi \, dm \). Then \( \mu : BC \to [0, \infty] \), \( \mu \) is linear, and \( \mu \) is positive (i.e. \( \phi(x) \geq 0 \) for all \( x \) implies \( \mu(\phi) \geq 0 \)).

If \( f : X \to X \) is continuous and sends bounded sets to bounded sets (e.g. if \( f \) is Lipschitz), then we define \( f_\# : \mathcal{M} \to \mathcal{M} \) by \( f_\#(\mu)(E) = \mu(f^{-1}(E)) \). Equivalently \( f_\#(\phi) = \mu(\phi \circ f) \). Notice that \( M(f_\#(\mu)) = M(\mu) \).


We define the weak topology on $\mathcal{M}$ by taking as a sub-basis all sets of the form $\{\mu : a < \mu(\phi) < b\}$, for arbitrary real $a < b$ and arbitrary $\phi \in \mathcal{BC}(X)$. It follows $\mu_{a} \to \mu$ in the weak topology if $\mu_{a}(\phi) \to \mu(\phi)$ for all $\phi \in \mathcal{BC}(X)$.

2.6. Hausdorff Measure.

(1) Let the real number $k \geq 0$ be fixed. For every $\delta > 0$ and $E \subset X$ we define

$$\mathcal{H}_{\delta}^{k}(E) = \inf \left\{ \sum_{i=1}^{\infty} \alpha_{k}2^{-k}(\text{diam } E_{i})^{k} : E \subset \bigcup_{i=1}^{\infty} E_{i}, \text{diam } E_{i} \leq \delta \right\}$$

$$\mathcal{H}^{k}(E) = \lim_{\delta \to 0} \mathcal{H}_{\delta}^{k}(E) = \sup_{\delta \geq 0} \mathcal{H}_{\delta}^{k}(E).$$

$\mathcal{H}^{k}(E)$ is called the Hausdorff $k$-dimensional measure of $E$. A reference is [FH, 2.10.3]. $\alpha_{k}$ is a suitable normalising constant. If $k$ is an integer, $\alpha_{k} = \mathcal{L}^{k}\{x \in \mathbb{R}^{k} : |x| \leq 1\}$. For arbitrary $k$ we define $\alpha_{k} = \Gamma(\frac{k}{2})/\Gamma(\frac{k}{2} + 1)$. The particular value of $\alpha_{k}$ for non-integer $k$ will not be important. The value of $\mathcal{H}^{k}(E)$, but not that of $\mathcal{H}_{\delta}^{k}(E)$, remains unchanged if we restrict the $E_{i}$ to be open (or closed, or convex).

$\mathcal{H}^{k}$ is a Borel regular measure, but $\mathcal{H}^{k}$ is not normally finite on bounded sets. If $X = \mathbb{R}^{n}$ then $\mathcal{H}^{n} = \mathcal{L}^{n}$, $\mathcal{H}^{0}$ is counting measure. If $f : A \subset \mathbb{R}^{m} \to \mathbb{R}^{n}$ is $C^{1}$ and one-one, then $\mathcal{H}^{m}(f(A)) = \int_{A} J(f) d\mathcal{L}^{m}$, where $J(f)$ is the Jacobian. Thus $\mathcal{H}^{k}$ agrees with usual notions of $k$-dimensional volume on "nice" sets in case $k$ is an integer.

If $F : X \to X$ is Lipschitz, then $\mathcal{H}^{k}(F(A)) \leq (\text{Lip } F)^{k}\mathcal{H}^{k}(A)$. If $F$ is a similitude, $F_{#}\mathcal{H}^{k} = (\text{Lip } F)^{-k}\mathcal{H}^{k}$.

For each $E \subset X$ there is a unique real number $k$, called the Hausdorff dimension of $E$, written $\dim E$, such that $\mathcal{H}^{\alpha}(E) = \infty$ if $\alpha < k$, $\mathcal{H}^{\alpha}(E) = 0$ if $\alpha > k$. $\mathcal{H}^{k}(E)$ can take any value in $[0, \infty]$.

(2) Suppose $S : X \to X$ is a similitude with $\text{Lip } S = r$. Then $\mathcal{H}^{k}(S(A)) = r^{k}\mathcal{H}^{k}(A)$. For $(\mathcal{H}^{k}(S(A))(E) = \mathcal{H}^{k}(S(A) \cap E) = \mathcal{H}^{k}(S(A \cap S^{-1}(E))) = r^{k}\mathcal{H}^{k}(A \cap S^{-1}(E)) = r^{k}(\mathcal{H}^{k}(A)(S^{-1}(E))) = r^{k}\mathcal{H}^{k}(A)(E)$.

(3) The lower (upper) $k$-dimensional density of the set $A$ at the point $x$ is defined respectively to be

$$\theta_{e}^{k}(A, x) = \lim_{r \to 0} \inf \frac{\mathcal{H}^{k}(A \cap B(x, r))}{\alpha_{k}r^{k}}.$$

$$\theta^{k}(A, x) = \lim_{r \to 0} \sup \frac{\mathcal{H}^{k}(A \cap B(x, r))}{\alpha_{k}r^{k}}.$$

If they are equal, their common value is called the $k$-dimensional density of $A$ at $x$, and is written $\theta^{k}(A, x)$.

Likewise, for $\mu$ a measure on $X$ we define

$$\theta_{e}^{k}(\mu, x) = \lim_{r \to 0} \inf \frac{\mu(B(x, r))}{\alpha_{k}r^{k}},$$

$$\theta^{k}(\mu, x) = \lim_{r \to 0} \sup \frac{\mu(B(x, r))}{\alpha_{k}r^{k}},$$

and $\theta^{k}(\mu, x)$ to be their common value if both are equal. Thus $\theta_{e}^{k}(A, x) = \theta^{k}(\mathcal{H}^{k}|A, x)$, and similarly for $\theta^{k}$, $\theta^{k}$.

Upper densities turn out to be more important than lower densities. The main results we will need are that for $\mu \in \mathcal{M}$,

(i) $\theta^{k}(\mu, a) \geq \lambda A$ for all $a \in A$ implies $\mathcal{H}^{k}(A) \leq \lambda^{-1}\mu(A)$,

(ii) $\theta^{k}(\mu, a) < \lambda A$ for all $a \in A$ implies $\mathcal{H}^{k}(A) \geq \lambda^{-1}\mu(A)$. 


In particular if $0 < \mu(A) < \infty$, and the upper density is bounded away from 0 and \( \infty \), this enables us to establish that $0 < \mathcal{H}^k(A) < \infty$. For a reference see [FH, 2.10.19 (1), (3)].

2.7. Geometric Measure Theory. We will briefly sketch the ideas from geometric measure theory needed for \S 6. A complete treatment is [FH], in particular Chapter 4, and a good exposition of the main results is in [FH1]. At a number of places we have found it convenient to abbreviate the standard notation.

1. Suppose $m \geq 0$ is a positive integer. A set $E \subset \mathbb{R}^n$ is \textit{m-rectifiable} iff $E$ is $\mathcal{H}^m$-measurable, $\mathcal{H}^m(E) < \infty$, and there exist $m$-dimensional $C^1$ manifolds $\{M_i\}_{i=1}^\infty$ in $\mathbb{R}^n$ such that $\mathcal{H}^m(E \sim \bigcup_{i=1}^\infty M_i) = 0$. (Here we differ somewhat from the convention of [FH]). For $\mathcal{H}^m$ a.a. $x \in E$ the tangent spaces at $x$ to distinct $M_i$ containing $x$ are equal. Let $\overline{E}_x$ be this tangent space where it exists.

2. Suppose now we are given

1. a bounded $m$-rectifiable set $E$ with $m \geq 1$,
2. a multiplicity function $\theta$, i.e. an $\mathcal{H}^m$-measurable function $\theta$ with domain $E$ and range a subset of the positive integers, such that $\int_E \theta \ d\mathcal{H}^m < \infty$,
3. an orientation $\overrightarrow{T}$, i.e. an $\mathcal{H}^m$-measurable function $\overrightarrow{T}$ with domain $E$ such that for $\mathcal{H}^m$ a.a. $x \in E$, $\overrightarrow{T}(x)$ is one of the two simple unit $m$-vectors associated with $\overline{E}_x$.

With the above ingredients we define a linear operator on $C^\infty$ $m$-forms $\phi$ by

$$T(\phi) = \int_E \theta(x) \langle \overrightarrow{T}(x), \phi(x) \rangle \ d\mathcal{H}^m.$$ 

This generalises the notion of integration over an oriented manifold. The set of all such operators is called the set of \textit{m-dimensional rectifiable currents}. A \textit{0-dimensional rectifiable current} is defined to be a linear operator $T$ on $C^\infty$ functions (i.e. $0$-forms) such that

$$T(\phi) = \sum_{i=1}^r \lambda_i \phi(a_i),$$

where $r \geq 0$, $\lambda_1, \ldots, \lambda_r$ are integers, and $a_1, \ldots, a_r \in \mathbb{R}^n$. Thus $T$ corresponds to a finite number of points with integer multiplicities. $T$ is written $\sum_{i=1}^r \lambda_i \lfloor[a_i]\rfloor$.

The set of $m$-dimensional rectifiable currents forms an abelian group in a natural way. It is denoted $\mathcal{R}_m$.

3. For each $T \in \mathcal{R}_m$, $m \geq 1$, we define a linear operator $\partial T$ on $C^\infty$ $(m-1)$-forms by Stokes formula:

$$\partial T(\phi) = T(d\phi).$$

If $T$ corresponds to a compact oriented manifold with boundary, then $\partial T$ corresponds to the oriented boundary. Clearly $\partial \partial T = 0$. However it is not necessarily true that $\partial T \in \mathcal{R}_{m-1}$. Accordingly we define the abelian group of $m$-dimensional \textit{integral currents} by

$$\mathcal{I}_m = \{ T \in \mathcal{R}_m : \partial T \in \mathcal{R}_{m-1} \} \quad \text{if } m \geq 0,$$

$$\mathcal{I}_0 = \mathcal{R}_0.$$

Clearly $\mathcal{I}_m \subset \mathcal{R}_m$. For $m \geq 1$, $\partial : \mathcal{I}_m \rightarrow \mathcal{I}_{m-1}$, and is a group homomorphism.

We can also enlarge $\mathcal{R}_m$ to the abelian group of $m$-dimensional \textit{integral flat chains}, or $m$-chains for short, defined by

$$\mathcal{F}_m = \{ R + \partial S : R \in \mathcal{R}_m, S \in \mathcal{R}_{m+1} \}.$$

In the natural way $\partial$ is extendible to a group homomorphism $\partial : \mathcal{F}_m \rightarrow \mathcal{F}_{m-1}$ if $m \geq 1$. 
(4) For $T \in \mathcal{R}_m$ we define the mass of $T$ by

$$M(T) = \int_E \theta \, d\mathcal{H}^m \quad \text{if } m \geq 1,$$

$$M\left(\sum_{i=1}^r \lambda_i[a_i]\right) = \sum_{i=1}^r |\lambda_i|.$$

One can extend the definition of $M$ to $\mathcal{F}_m$, but then one has [FH, 4.2.16].

$$\mathcal{R}_m = \mathcal{F}_m \cap \{T : M(T) < \infty\},$$

$$\mathcal{I}_m = \mathcal{R}_m \cap \{T : M(\partial T) < \infty\}.$$

One now defines the integral flat “norm” on $\mathcal{F}_m$ by

$$\mathcal{F}(T) = \inf \{M(R) + M(S) : T = R + \partial S\},$$

and the integral flat metric by

$$\mathcal{F}(T_1, T_2) = \mathcal{F}(T_1 - T_2).$$

![Figure 2.1](image)

Thus $T_1$ and $T_2$ of Figure 2.1 are close in the $\mathcal{F}$-metric since there exist $R$ and $S$ of small mass such that $T_1 - T_2 = R + \partial S$. $\partial$ is continuous in the $\mathcal{F}$-metric, indeed $\mathcal{F}(\partial T_1, \partial T_2) \leq \mathcal{F}(T_1, T_2)$.

(5) The $m$-dimensional integral flat chains generalise the notion of an oriented $m$-dimensional $C^1$ manifold, but retain many of the desirable properties and at the same time are closed under various useful operations.

Thus $\mathcal{F}_m$ is a complete metric space under $\mathcal{F}$ [FH, 4.1.24]. The infimum in the definition of $\mathcal{F}$ is always realised [FH, 4.2.18].

Convergence in $M$ implies convergence in $\mathcal{F}$, but certainly not conversely. If $T_j \to T$ in $\mathcal{F}$, then $M(T) \leq \liminf M(T_j)$.

For integral currents there is the important compactness theorem [FH, 4.2.17]: if $K \subset \mathbb{R}^n$ is compact and $c < \infty$, then

$$\{T \in \mathcal{I}_m : M(T) < c, \ M(\partial T) < c, \ \text{spt } T \subset K\}$$

is compact in the $\mathcal{F}$-topology. For $T \in \mathcal{F}_m$ we define $\text{spt } T$, the support of $T$, to be the intersection of all closed sets $C$ such that $\text{spt } \phi \cap C = \emptyset$ implies $T(\phi) = 0$.

If $T \in \mathcal{F}_m$, $m \geq 1$, and $\partial T = 0$ (or if $T \in \mathcal{F}_0$), we say $T$ is an $m$-dimensional integral flat cycle or $m$-cycle for short. If $m \geq 1$, it follows by a cone construction [FH, 4.1.11] that $T = \partial S$ for some $S \in \mathcal{F}_{m+1}$. Furthermore, one has the isoperimetric inequality [FH, 4.2.10]: for $m \geq 1$ there is a constant $\gamma = \gamma(m, n)$ depending only on $m$ and $n$, such that if $T \in \mathcal{I}_m$ and $\partial T = 0$, then $T = \partial S$ for some $S \in \mathcal{I}_{m+1}$ with $M(S) \leq \gamma M(T)^{m+1/n}$.

If $T \in \mathcal{F}_m$ and $T = \partial S$ for some $S \in \mathcal{F}_{m+1}$, we say $T$ is an $m$-dimensional integral flat boundary, or $m$-boundary for short. Thus if $m \geq 1$, every $m$-cycle is an $m$-boundary.

(6) If $T \in \mathcal{F}_m$ and $f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz and proper, then one defines $f_T \in \mathcal{F}_m$ [FH, 4.1.14, 4.1.24]. In case $T$ corresponds to an oriented manifold and
f is $C^1$, then $f_\#T$ corresponds to the oriented image of $T$ under $f$, with appropriate multiplicities if $f$ is not one-one.

The properties of $f_\#T$ we will need are:

(a) $f_\# \partial T = \partial f_\#T$;
(b) $f : \mathcal{F}_m \to \mathcal{F}_m$, is linear, and is continuous in the $\mathcal{F}$-metric;
(c) if $\text{Lip} f = r$ and $T \in \mathcal{F}_m$, then $M(f_\#T) \leq r^m M(T)$ and $\mathcal{F}(f_\#T) \leq \max\{r^m, r^{m+1}\} \mathcal{F}(T)$;
(d) $f_\#T \subset f(\text{spt} T)$.

(7) One can generalise from the integral flat chains to the so-called flat chains, and even more generally to the currents of de Rham. However, one loses the useful geometric properties of the integral flat chains. For a full treatment of all these subjects see [FH].

3. IN Variant Sets

We follow the notation of 2.2, and the other subsections of 2 as necessary. $S = \{S_1, \ldots, S_N\}$ is a set of contraction maps on the complete metric space $(X, d)$.

Lip $S_i = r_i$. $s_i, \ldots, i_p$ is the fixed point of $S_{i_1 \ldots i_p}$.

We show the existence and uniqueness of a compact set invariant with respect to $S$, and discuss its properties.

We suggest the reader considers Examples 3.3 for motivation.


(1) For arbitrary $A \subset X$ let $S(A) = \bigcup_{i=1}^N S_i(A)$. Let $S^0(A) = A$, $S^1(A) = S(A)$, $S^p(A) = S(S^{p-1}(A))$ for $p \geq 2$. We will often use the notation $A_{i_1 \ldots i_p} = S_{i_1 \ldots i_p}(A)$. Notice $S^0(A) = \bigcup_{i_1 \ldots i_p} A_{i_1 \ldots i_p}$. Notice also that $\text{diam} (A_{i_1 \ldots i_p}) \leq r_{i_1} \cdots r_{i_p}$ diam $(A) \to 0$ as $p \to \infty$, provided $A$ is bounded.

(2) Definition. A is invariant (with respect to $S$) if $A = S(A)$.

3.2. Theorem and Definitions.

(i) There is a unique closed bounded set $K$ which is invariant with respect to $S$. Thus $K = \bigcap_{k=1}^N K_i$. Moreover $K$ is compact.

(ii) $K_{i_1 \ldots i_p} = \bigcup_{p+1}^N K_{i_1 \ldots i_p+1}$.

(iii) $K \supseteq K_{i_1} \supseteq \cdots \supseteq K_{i_1 \ldots i_p} \supseteq \cdots$, and $\bigcap_{i=1}^\infty K_{i_1 \ldots i_p}$ is a singleton whose member is denoted $k_{i_1 \ldots i_p}$. $K$ is the union of these singletons.

(iv) $k_{i_1 \ldots i_p} = s_{i_1 \ldots i_p}$, and in particular $s_{i_1 \ldots i_p} \in K$ (recall 2.1(5)). Also $k_{i_1 \ldots i_p} = \lim_{p \to \infty} s_{i_1 \ldots i_p}$, and in particular this limit exists.

(v) $K$ is the closure of the set of fixed points of the $S_{i_1 \ldots i_p}$.

(vi) $S_{j_1 \ldots j_p}(K_{i_1 \ldots i_p}) = K_{j_1 \ldots j_{p+1} i_1 \ldots i_p}$, $S_{j_1 \ldots j_p}(k_{i_1 \ldots i_p}) = k_{j_1 \ldots j_{p+1} i_1 \ldots i_p}$.

(vii) The coordinate map $\pi : C(N) \to K$ given by $\pi(\alpha) = k_\alpha$ is a continuous map onto $K$.

(viii) If $A$ is a non-empty bounded set, then $d(A_{i_1 \ldots i_p}, k_{i_1 \ldots i_p}) \to 0$ uniformly as $p \to \infty$. In particular $S^p(A) \to K$ in the Hausdorff metric.

(4) Proof of Uniqueness. We first remark that (i) and (viii) are established in 3.2 independently of the following.
Assume now that $K$ is a closed bounded set invariant with respect to $S$, and observe the following consequences.

$$K = \bigcup_{i=1}^{N} S_i(K) = \bigcup_{i,j} S_i(S_j(K)) = \bigcup_{i,j} S_{ij}(K) = \bigcup_{i,j} K_{ij}$$

$$\ldots = \bigcup_{i_1, \ldots, i_p} K_{i_1 \ldots i_p}.$$  

Similarly

$$K_{i_1 \ldots i_p} = S_{i_1 \ldots i_p}(K) = S_{i_1 \ldots i_p} \left( \bigcup_{i_{p+1}=1}^{N} S_{i_{p+1}}(K) \right)$$

$$= \bigcup_{i_{p+1}=1}^{N} S_{i_1 \ldots i_p i_{p+1}}(K) = \bigcup_{i_{p+1}=1}^{N} K_{i_1 \ldots i_p i_{p+1}}.$$  

Thus $K \supseteq K_{i_1} \supseteq K_{i_1 i_2} \supseteq \ldots \supseteq K_{i_1 \ldots i_p} \supseteq \ldots$, and since diam $(K_{i_1 \ldots i_p})$ as $p \to \infty$, $\bigcap_{p} K_{i_1 \ldots i_p}$ is a singleton (by completeness of $X$) whose unique member we denote $k_{i_1 \ldots i_p}$. Thus we have established $(ii)$ and $(iii)$ under the given assumptions on $K$.

The first part of $(vi)$ is immediate, since

$$S_{i_1 \ldots i_p}(K_{i_1 \ldots i_p}) = S_{i_1 \ldots i_p} \left( S_{i_1 \ldots i_p}(K) \right) = S_{i_1 \ldots i_p i_1 \ldots i_p}(K) = K_{i_1 \ldots i_p i_1 \ldots i_p}.$$  

The second part follows, since

$$S_{i_1 \ldots i_p}(k_{i_1 \ldots i_p}) \in S_{i_1 \ldots i_p} \bigcap_{p=1}^{\infty} K_{i_1 \ldots i_p} = \bigcap_{p=1}^{\infty} K_{i_1 \ldots i_p i_1 \ldots i_p} = k_{i_1 \ldots i_p i_1 \ldots i_p}.$$  

Since $S_{i_1 \ldots i_p}(k_{i_1 \ldots i_p}) = k_{i_1 \ldots i_p}$ by the above, it follows $k_{i_1 \ldots i_p}$ is the unique fixed point $s_{i_1 \ldots i_p}$ of $S_{i_1 \ldots i_p}$. It follows both $s_{i_1 \ldots i_p}$, $k_{i_1 \ldots i_p} \in K_{i_1 \ldots i_p}$, and hence since $\lim_{p \to \infty} \text{diam } (K_{i_1 \ldots i_p}) = 0$, that $\lim_{p \to \infty} S_{i_1 \ldots i_p} = k_{i_1 \ldots i_p}$. This establishes $(iv)$, and $(v)$ follows from $(iv)$. Notice we have established the uniqueness of $K$ (since $K$ is the union of singletons, each of which is the limit of a certain sequence of fixed points of the $S_{i_1 \ldots i_p}$).

To establish $(vii)$, and hence that $K$ is compact (being the continuous image of a compact set), let $\pi$ be as in $(vii)$. Suppose $\alpha = (\alpha_1, \ldots, \alpha_p, \ldots) \in C(N)$ and $\varepsilon > 0$. Then $\pi(\alpha) = k_{\alpha_1, \alpha_2, \ldots}$ and so there is a $q$ such that $K_{\alpha_1, \ldots, \alpha_q} \subseteq \{ x \in K : d(x, \pi(\alpha)) < \varepsilon \}$. Since $K_{\alpha_1, \ldots, \alpha_q}$ is the image of the open set $\{ \beta : \beta_i = \alpha_i \text{ if } i \leq q \}$, it follows $\pi$ is continuous.

To prove $(viii)$ suppose $A$ is non-empty and bounded. Then

$$d(A_{i_1 \ldots i_p}, k_{i_1 \ldots i_p}) = d(S_{i_1 \ldots i_p}(A), S_{i_1 \ldots i_p}(k_{i_{p+1}}))$$

$$\leq r_{i_1} \cdots r_{i_p} d(A, k_{i_{p+1}})$$

$$\leq r_{i_1} \cdots r_{i_p} \sup \{ d(a, b) : a \in A, b \in K \}$$

$$\leq \text{Constant} (\max_{1 \leq i \leq N} r_i)^p$$

$$\to 0 \quad \text{as } p \to \infty.$$  

All that remains now is to prove the existence of a closed bounded invariant set. But notice that we know from $(v)$ what this set must be.

5 Proof of Existence. First we need to establish the following lemma.

Lemma. If $\{ S_1, \ldots, S_N \}$ is a set of contraction maps on a complete metric $(X, d)$, and $s_{i_1 \ldots i_p}$ is the fixed point of $S_{i_1 \ldots i_p} = S_{i_1} \circ \cdots \circ S_{i_p}$ then for each sequence $i_1 \ldots i_p \ldots$, $\lim_{p \to \infty} S_{i_1 \ldots i_p}$ exists.
Let $\lambda = \max_{1 \leq i, j \leq N} d(s_i, s_j)$, and let $R = \lambda (1-r)^{-1}$ where $r = \max \{r_i = \text{Lip}(S_i) : 1 \leq i \leq N \}$.

Then $\bigcap_{i=1}^{N} B(s_i, rR) \subset \bigcap_{i=1}^{N} B(s_i, R) = C$, say. For if $d(s_i, x) \leq rR$ then $d(s_j, x) \leq \lambda + rR = \lambda + r(1-r)^{-1} = \lambda (1-r)^{-1} = R$. Thus $S_i C \subset C$ for $i = 1, \ldots, N$, and so $C \supset S_1(C) \supset S_{1, 1}(C) \supset \cdots \supset S_{1, \ldots, p}(C) \supset \cdots$, i.e. $C \supset C_1 \supset C_{1, 1} \supset \cdots \supset C_{1, \ldots, p} \supset \cdots$. But the fixed point $s_{1, \ldots, p}$ must lie in $S_{1, \ldots, p}(C)$, and so since $\text{diam}(S_{1, \ldots, p}(C)) = 0$ as $p \to \infty$ and the $S_{1, \ldots, p}(C)$ are closed, it follows $\lim_{p \to \infty} s_{1, \ldots, p}$ exist and is the unique member of $\bigcap_{p=1}^{\infty} S_{1, \ldots, p}(C)$.

For $\alpha \in C(N)$ let $s_\alpha = \lim_{p \to \infty} s_{\alpha_1, \ldots, \alpha_p}$, and let $K = \{s_\alpha : \alpha \in C(N)\}$. Then $S_i(s_\alpha) = s_\alpha$, since $S_i(s_\alpha) \in S_i(\bigcap_{p=1}^{\infty} C_{\alpha_1, \ldots, \alpha_p}) = \bigcap_{p=1}^{\infty} C_{\alpha_1, \ldots, \alpha_p} \supset s_\alpha$ (notice that it is not normally true that $S_i(s_{\alpha_1, \ldots, \alpha_p}) = (s_{\alpha_1}, \ldots, \alpha_p)$). Thus $K = \bigcup_{i=1}^{\infty} S_i(K) = S(K)$, i.e. $K$ is invariant with respect to $S$.

It remains to prove that $K$ is compact. Define $\pi : C(N) \to K$ by $\pi(\alpha) = s_\alpha$. Since $\text{diam}(K)$ is bounded (being a subset of $C$ in the previous lemma) it follows precisely as in the proof of (vii) that $\pi$ is continuous and hence that $K$ is compact. This gives the existence of (i) and completes the proof of the theorem. \[\Box\]

6) Definition. The compact set invariant under $S$ is denoted by $|S|$.

7) Non-compact invariant sets. There are always non-bounded invariant sets, $\mathbb{R}^n$ being a trivial example.

For any $A$, $S(A) = A$ implies $S(\bar{A}) = \bar{A}$. Thus if $A$ is bounded and invariant, then so is $A$, and hence $\bar{A} = |S|$ by (3)(i). For example, $S(0, 1) = (0, 1)$ where $S(x)$ is as in 3.1(1).

8) The following observation is useful. Suppose $A$ is a set such that $S(A) \subset A$. Then clearly $A \supset S(A) \supset S^2(A) \supset \cdots \supset S^n(A) \supset \cdots$. If furthermore $A$ is closed and non-empty, then $K = |S| \subset A$, and $K_{i_1, \ldots, i_p} \subset A_{i_1, \ldots, i_p}$ for all $i_1, \ldots, i_p$.

To see this latter, choose $a \in A$. Then by (3)(viii) for each fixed $i_1, \ldots, i_p, \ldots$, $k_{i_1, \ldots, i_p} = \lim_{p \to \infty} S_{i_1, \ldots, i_p}(a) \in A$. Hence $K \subset A$. Applying $S_{i_1, \ldots, i_p}$ to both sides, $K_{i_1, \ldots, i_p} \subset A_{i_1, \ldots, i_p}$.

9) If $\sum_{p=1}^{N} r_i < 1$, then $K$ is totally disconnected. For given $a, b \in K$ select $p$ such that $\lambda \left( \sum_{i_1, \ldots, i_p} r_i \right)^p = \lambda \left( \sum_{i_1, \ldots, i_p} r_i \right)^p < d(a, b)$, where $\lambda = \text{diam}K$. Since $K = \bigcup_{i_1, \ldots, i_p} K_{i_1, \ldots, i_p}$, and $\text{diam}K_{i_1, \ldots, i_p} = r_i \cdots r_p \lambda$, it follows by an elementary argument that $a$ and $b$ are in distinct components of $K$.

3.2. Convergence in the Hausdorff Metric. We remark that this Section is independent of 3.1(3), (4).

Let $B$ be the family of closed bounded subsets of $X, C$ the family of compact subsets. Clearly $S : B \to B$ and $S : C \to C$. We have

1) Theorem. $S$ is a contraction map on $B$ (respectively $C$) in the Hausdorff metric.

Proof.

\[\delta(S(A), S(B)) = \delta \left( \bigcup_i S_i(A), \bigcup_i S_i(B) \right) \leq \max_{1 \leq i \leq N} \delta(S_i(A), S_i(B)) \leq (\max_{1 \leq i \leq N} r_i) \delta(A, B).\]

Existence and uniqueness of a closed bounded invariant set $|S|$ follow from the contraction mapping principle. Since $C$ is a closed subset of $B$, it follows that $|S| \in C$. \[\Box\]
Remark added subsequent to publication: As pointed out by a number of people, since $A \subset B$ does not imply $S_i(A)$ closed, one should replace $S_i(A)$ by its closure in the definition of $S(A)$, when $S$ operates on $B$. This new map $S$ has a unique fixed point in $B$, which must then agree by uniqueness with the fixed point of $S$ operating on $C$.

3.3. Examples.

(1) **Cantor Set.** In the notation of 2.3 let

$$S_r = \{S_1(r), S_2(r)\}, \quad S_i(r) : \mathbb{R} \to \mathbb{R},$$

$$S_1(r) = (0, r, I), \quad S_2(r) = (1, r, I),$$

where $I$ is the identity map.

![Figure 3.1](image1) The classical Cantor set $C$.

If $r = \frac{1}{3}$, then $S_r(C) = C$ where $C$ is the classical Cantor set, and so $|S_{\frac{1}{3}}| = C$. We have sketched $C$, more precisely $S_{\frac{1}{3}}([0, 1])$, in Figure 3.1. Notice the numbering system for the various components $C_{1, 2, \ldots}$.

If $0 < r < \frac{1}{2}$, then $|S_r|$ is a generalised Cantor set. It is standard, and a consequence of 5.3(1)(ii), that $\dim |S_r| = \log 2/\log(\frac{1}{r})$.

If $\frac{1}{2} \leq r < 1$, then $S_r([0, 1]) = [0, 1]$, and hence $|S_r| = [0, 1]$. Thus different $S_r$ can generate the same set. In this connection see 4.1.

(2) **Koch Curve.** We refer to Figure 3.2. Let $a_1, a_2, a_3, a_4, a_5$ be as shown. Let $S = \{S_1, S_2, S_3, S_4\}$ where $S_i : \mathbb{R}^2 \to \mathbb{R}^2$ is the unique similitude mapping $\vec{a_i} \to \vec{a_i} + \vec{s_i}$ to $a_i a_{i+1}$ and having positive determinant (i.e. no reflections).

![Figure 3.2](image2) The Koch curve $K$.

Let $K = |S|$. Actually Figure 3.2 shows the approximation $S^4([a_1, a_5])$ to $K$. Notice how one finds the components of $K$, e.g. $K_{331}$. $S_i$ has the fixed point $s_i = k_i$.
\[ s_1 = a_1, \quad s_4 = a_5, \quad \text{and} \quad s_2, s_3 \text{ are shown. Similarly } S_{ij} = S_i \circ S_j \text{ has the fixed point} \quad s_{ij} = k_{ij}, \quad \text{where } s_{23} \text{ is shown.} \]

One can visualise \( \mathcal{S}(A) = \bigcup_{i_1, \ldots, i_p} A_{i_1 \ldots i_p} \to K \) as \( p \to \infty \), for arbitrary bounded \( A \) (e.g. \( A \) a singleton).

Now let \( S' = \{ S'_1, S'_2 \} \), where \( S'_i \) is the unique similitude mapping \( \overline{a_i a_3} \) to \( \overline{a_i a_3} \) (\( i = 1 \)), \( a_3 a_\delta \) (\( i = 2 \)), having negative determinant (i.e. the \( S'_i \) include a reflection component). Then it follows \( S'(K) = K \) and hence \( |S'| = K \). For \( S'_1 \circ S'_1 = S_1, \) \( S'_1 \circ S'_2 = S_2, \) \( S'_1 \circ S'_1 = S_3, \) \( S'_2 \circ S'_2 = S_4, \) hence \( (S')^2 = S \), hence \( |S'| \) is fixed by \( S \), hence \( |S'| = |S| \) by uniqueness. Thus as in (2), different \( S \) can generate the same set.

(3) Let \( M \subset \mathbb{R}^n \) be an oriented \( m \)-dimensional manifold with oriented boundary \( N \) as in Figure 3.3. Let \( \mathcal{S} = \{ S_1, \ldots, S_N \} \) where \( S_i : \mathbb{R}^n \to \mathbb{R}^n \) are contraction maps such that \( \sum_{i=1}^N S_i(N) = N, \) taking into account orientation and after allowing cancellation of portions of manifolds having opposite orientation. Obviously such \( \mathcal{S} \) are easy to find.

![Figure 3.3](image)

\[ M \subset \mathbb{R}^3, \] \( M \) has the boundary \( N. \) Consider the case where \( M \) and \( N \) do not lie in a plane.

\[ |\mathcal{S}| \] will normally have dimension \( > m, \) and so cannot be an \( m \)-dimensional manifold, yet in some sense \( |\mathcal{S}| \) is an \( m \)-dimensional object with oriented boundary \( N. \) We make this precise in §6, where under mild restrictions on the \( S_i, \) \( |\mathcal{S}| \) becomes an integral flat \( m \)-chain having \( N \) as its boundary.

3.4. Remark. The following gives a curious characterisation of line segments:

A compact connected set \( A \subset \mathbb{R}^n \) is a line segment iff \( A = S(A) \) for some \( \mathcal{S} = \{ S_1, \ldots, S_N \} \) where \( N \geq 2, \) the \( S_i \) are similitudes, \( \text{Lip} S_i = r_i, \) and \( \sum_{i=1}^N r_i = 1. \)

Proof. One direction is trivial.

Conversely, suppose \( A = S(A) \) with \( \mathcal{S} \) as above. Let \( \text{diam} A = d(p, q) \) where \( p, q \in A. \) By projecting \( A \) onto the line segment \( \overline{pq}, \) one sees that \( \mathcal{H}^1(A) \geq \text{diam} A. \) If \( A \neq \overline{pq}, \) by taking the nearest point retraction \( \pi \) of \( A \) onto a suitably thin solid ellipsoid having \( p \) and \( q \) as extremal points, one finds \( \mathcal{H}^1(A) \gtrless \mathcal{H}^1(\pi(A)). \) But \( \mathcal{H}^1(\pi(A)) \geq \text{diam} A \) as we just saw, and so \( \mathcal{H}^1(A) \gtrless \text{diam} A \) unless \( A = \overline{pq}. \) One can check \( \mathcal{H}^1(A) \leq \text{diam} A \) by a covering argument, and so the required result follows.

The above was in response to a query of B. Mandelbrot concerning characterisations of the line — his query in turn arose from some rather vague remarks in a work of Liebniz. F. J. Almgren Jr. suggested a shortening of the original proof.
3.5. Parametrised Curves.

(1) Suppose \( S = \{S_1, \ldots, S_N\} \) has the property that
\[
\begin{align*}
a &= s_1 = \text{fixed point of } S_1, \\
b &= s_N = \text{fixed point of } S_N, \\
S_i(b) &= S_{i+1}(a) \quad \text{if } 1 \leq i \leq N - 1,
\end{align*}
\]
(by example, 3.3(2)). Then one can define a continuous \( f : [0, 1] \rightarrow |S| \), with Image \((f) = |S|\), in a natural way.

For this purpose, fix \( 0 = t_1 < t_2 < \cdots < t_{N+1} = 1 \). Define \( g_i : [t_i, t_{i+1}] \rightarrow (0, 1) \) for \( 1 \leq i \leq N \) by
\[
g_i(x) = \frac{x - t_i}{t_{i+1} - t_i}.
\]

Let
\[
\mathcal{F} = \mathcal{F}(a, b) = \{f : [0, 1] \rightarrow X : f \text{ is continuous, } f(0) = a, f(1) = b\}.
\]

Define \( S(f) \) for \( f \in \mathcal{F} \) by
\[
S(f)(x) = S_i \circ f \circ g_i(x) \quad \text{for } x \in [t_i, t_{i+1}], \ 1 \leq i \leq N.
\]

Define a metric \( \mathcal{P} \) on \( \mathcal{F} \) by
\[
\mathcal{P}(f_1, f_2) = \sup\{|f_1(x) - f_2(x)| : x \in [0, 1]\}.
\]

\( \mathcal{P} \) is clearly a metric, and it is furthermore complete since the uniform limit of continuous functions is continuous.

(2) Proposition. \( S \) is well-defined, \( S : \mathcal{F} \rightarrow \mathcal{F}, \) and \( S \) is a contraction map in the metric \( \mathcal{P} \).

Proof. \( S \) is well-defined and \( S(f) \in \mathcal{F} \) if \( f \in \mathcal{F} \), since \( S_i \circ f \circ g_i(t_{i+1}) = S_i \circ f(1) = S_i(b) = S_{i+1}(a) = S_{i+1} \circ f(0) = S_{i+1} \circ f \circ g_{i+1}(t_{i+1}) \) for \( 1 \leq i \leq N - 1 \), \( S_1 \circ f \circ g_1(0) = S_1 \circ f(0) = S_1(a) = a \), and \( S_N \circ f \circ g_N(1) = S_N \circ f(1) = S_N(b) = b \).

Now suppose \( x \in [t_i, t_{i+1}] \) and \( f_1, f_2 \in \mathcal{F} \). Then
\[
|S(f_1)(x) - S(f_2)(x)| = |S_i \circ f \circ g_i(x) - S_i \circ f_2 \circ g_i(x)|
\]
\[
\leq \text{Lip}\, S_i |f_1(g_i(x)) - f_2(g_i(x))| \\
\leq \text{Lip}\, S_i \mathcal{P}(f_1, f_2).
\]

Hence \( \mathcal{P}(S(f_1), S(f_2)) \leq r \mathcal{P}(f_1, f_2) \), where \( r = \max\{\text{Lip}\, S_i : 1 \leq i \leq N\} \). It follows \( S \) is a contraction map. \( \square \)

(3) Theorem. Under the hypotheses on \( S \) in (1), there is a unique \( g \in \mathcal{F} \) such that \( S(g) = g \). Furthermore Image \((g) = |S|\).

Proof. The existence of a unique such \( g \) follows from (2).

By construction, Image \( S(f) = S(\text{Image} \, f) \) for every \( f \in \mathcal{F} \). If \( S(g) = g \), this implies Image \( g = S(\text{Image} \, g) \), and hence Image \( g \equiv |S| \) by 3.1(3)(i). \( \square \)

(4) It is often possible to parametrise other invariant sets \(|S| \subset \mathbb{R} \) by maps \( g : \{x \subset \mathbb{R}^n : |x| \leq 1\} \rightarrow \mathbb{R}^n \) for suitable \( m \), for example \( m = 2 \) in 3.3(3). But if \( m > 1 \) there is a lot of arbitrariness in the selection of the particular parametric map \( g \). It is often better to treat \(|S|\) as an intrinsic “\( m \)-dimensional” object in \( \mathbb{R}^n \) via the notion of an \( m \)-dimensional integral flat chain, c.f. 2.7, and 6.
4. Invariant Measures

4.1. Motivation. A motivation for this section is the following. In 3.3(1), (2) we saw examples of different families of contractions generating the same set. Yet the $S_i$ of 3.3(1) seem to be different from one another in a way that $S$ and $S'$ of 3.3(2) are not. We make this precise in 4.4(6).

Another motivation is that it will be easier to "use" invariant sets if we can impose additional natural structure on them, in this case a measure.

4.2. Definitions. $(X, d)$ is a complete metric space, $S = \{S_1, \ldots, S_N\}$ is a family of contraction maps. Additionally, we assume the existence of a set $\rho = \{\rho_1, \ldots, \rho_N\}$ with $\rho_i \in (0,1)$ and $\sum_{i=1}^N \rho_i = 1$. In 5 we will see that in case the $S_i$ are similitudes with $\text{Lip } S_i = r_i$, it is natural to take $\rho_i = r_i^D$, where $D$ is the similarity dimension of $S$, 5.1(3).

We refer back to 2.5 for terminology on measure.

(1) Definition. If $\nu \in \mathcal{M}$ let $(S, \rho)(\nu) = \sum_{i=1}^N \rho_i S_i \# \nu$. Thus $(S, \rho)(\nu)(A) = \sum_{i=1}^N \rho_i (S_i^{-1}(A))$. Let $(S, \rho)^0(\nu) = \nu$, $(S, \rho)^1(\nu) = (S, \rho)(\nu)$, and $(S, \rho)^p(\nu) = (S, \rho)((S, \rho)^{p-1}(\nu))$ for $p \geq 2$. Let

$$\nu_{i_1 \ldots i_p} = \rho_{i_1} \cdot \ldots \cdot \rho_{i_p} S_{i_1 \ldots i_p} \# (\nu).$$

(2) Notice $(S, \rho)^p(\nu) = \sum_{i_1 \ldots i_p} \nu_{i_1 \ldots i_p}$. Also $M((S, \rho)(\nu)) = M(\nu)$ and so $M((S, \rho)^p(\nu)) = M(\nu)$ for all $p$. In particular, $(S, \rho) : \mathcal{M}^1 \to \mathcal{M}^1$.

(3) Definition. $\nu$ is invariant (with respect to $(S, \rho)$) if $(S, \rho)(\nu) = \nu$.

4.3. The L metric. We introduce a metric $L$ on $\mathcal{M}^1$ (see 2.5(2)) similar to the one introduced by Almgren in [AF, 2.6], but modified in a way which enables 4.4(1) to hold.

(1) Definition. For $\mu, \nu \in \mathcal{M}^1$ let

$$L(\mu, \nu) = \sup \{\mu(\phi) - \nu(\phi) \mid \phi : X \to R, \text{Lip } \phi \leq 1\}.$$ 

Notice that $\phi$ of the definition is a member of $BC(X)$, and notice also that there is no restriction on $\sup \{\phi(x) \mid x \in X\}$.

In checking that $L$ is indeed a metric, the only part which is not completely straightforward is verifying $L(\mu, \nu) < \infty$. So suppose $\mu \cup \text{spt } \nu \subset B(a,R)$. Then for $\text{Lip } \phi \leq 1$,

$$\mu(\phi) - \nu(\phi) = \mu(\phi - \phi(a) + \phi(a)) - \nu(\phi - \phi(a) + \phi(a))$$

$$= \mu(\phi - \phi(a)) - \nu(\phi - \phi(a)), \quad \text{since } \mu(\phi(a)) = \nu(\phi(a))$$

$$\leq \mu(R) + \nu(R)$$

$$= 2R.$$

One can check that the $L$ metric topology and the weak topology coincide on $H^1 \cap \{\mu : \text{spt } \mu \text{ is compact}\}$.

Finally notice that $L(\delta_a, \delta_b) = d(a,b)$.

4.4. Existence and Uniqueness.

(1) Theorem.

(i) $(S, \rho) : \mathcal{M}^1 \to \mathcal{M}^1$ is a contraction map in the $L$ metric.

(ii) There exists a unique $\mu \in \mathcal{M}^1$ such that $(S, \rho)\mu = \mu$. If $\nu \in \mathcal{M}^1$ then $(S, \rho)^p(\nu) \to \mu$ is the $L$ metric, and hence in the topology of convergence with respect to each compactly supported continuous function.
Proof. (ii) follows immediately from (i).

To establish (i), suppose Lip $\phi \leq 1$ and let $r = \max_{1 \leq i \leq N} r_i$. Then for $\mu, \nu \in \mathcal{M}^1$,

$$
(S, \rho)(\mu)(\phi) - (S, \rho)(\nu)(\phi) = \sum_{i=1}^{N} (\rho_i S_i \# \mu)(\phi) - \sum_{i=1}^{N} (\rho_i S_i \# \nu)(\phi)
$$

$$
= \sum_{i=1}^{N} \rho_i (\phi \circ S_i - \nu(\phi \circ S_i))
$$

$$
= \sum_{i=1}^{N} \rho_i r (\mu(r^{-1} \phi \circ S_i) - \nu(r^{-1} \phi \circ S_i))
$$

$$
\leq \sum_{i=1}^{N} \rho_i r L(\mu, \nu) = r L(\mu, \nu),
$$

since Lip $(r^{-1} \phi \circ S_i) \leq r^{-1} \cdot 1 \cdot r_i \leq 1$. \hfill \Box

(2) Definition. The unique measure invariant with respect to $(S, \rho)$ is denoted $\|S, \rho\|$.

(3) Definition. Let $\tau$ be the product measure on $\mathcal{C}(N)$ induced by the measure $\rho(i) = \rho_i$ on each factor $\{1, \ldots, N\}$.

(4) Theorem.

(i) $\|S, \rho\| = \pi_\# \tau$, where $\pi : \mathcal{C}(N) \to K$ is the coordinate map of 3.1(3)(vii).

(ii) $\text{spt } \|S, \rho\| = |S|$.

Proof. (ii) follows from (i) and 3.1(3)(vii).

To establish (i) let $\sigma_i : \mathcal{C}(N) \to \mathcal{C}(N)$ be the $i$th shift operator 2.1(3). Clearly $\pi \circ \sigma_i = S_i \circ \pi$ and $\tau$ is $(\{\sigma_1, \ldots, \sigma_N\}, \rho)$ invariant. Hence $\sum \rho_i S_i \# (\pi_\# \tau) = \sum \rho_i S_i \# (\sigma_i \# \tau) = \pi_\# \sum \rho_i (\sigma_i \# \tau) = \pi_\# \tau$, and so by uniqueness $\pi_\# \tau = \|S, \rho\|$. \hfill \Box

(5) Remarks.

(1) It follows from 4(ii) that $\|S, \rho\|$ has compact support.

(2) There are unbounded invariant measures, in particular and trivially, $\mathcal{L}^n$ on $\mathbb{R}^n$.

(3) If $\nu$ is invariant, so is $\lambda \nu$ for any positive constant $\lambda$. Requiring $\|S, \rho\| \in \mathcal{M}^1$ is simply a normalisation requirement.

(6) Example. Referring back to 3.3(1), let $\rho = (\frac{1}{2}, \frac{1}{2})$ and write $\mu_r$ for $\|S, \rho\|$. We will show $\mu_r \neq \mu_s$ for $\frac{1}{2} \leq r < s < 1$. In 5.3(iii) we see that $\mu_r = \mathcal{H}^r(\|S, \rho\|)$ for $0 < r \leq \frac{1}{2}$.

Suppose $\frac{1}{2} \leq r < s < 1$. Take $A \subset [0, 1 - s)$. Then $S_2(r^{-1})(A) \cap [0, 1] = \emptyset$ and hence $\mu_r(S_2(r^{-1})(A)) = 0$ since $\text{spt } \mu_r \subset [0, 1]$. It follows $\mu_r(A) = \frac{1}{2} \mu_r(S_2(r^{-1})(A)) = \frac{1}{2} \mu_r(rA)$. Similarly $\mu_s(A) = \frac{1}{2} \mu_s(sA)$. If $\mu_r = \mu_s = \mu$, say, it follows $\mu(sA \sim rA) = 0$. Choosing $A = [0, 1 - s)$, this contradicts $\text{spt } \mu = [0, 1]$.

4.5 Different Sets of Similitudes Generating the Same Set. For $I$ a finite set as in 2.1(6), let $S_I = \{S_\alpha : \alpha \in I\}$. Then $|S_I| = \{k_\beta : \beta \in \hat{I}\}$, as follows by applying 3.3(3)(iii), (iv) to $S_I$. From 2.1(8)(i), $I = I$ iff $I$ is secure. Thus $|S_I| = |S|$ if $I$ is secure, and if the coordinate map $\pi$ is one-one, then $|S_I| = |S|$ iff $I$ is secure. A similar result was first shown in [OP].

Let $\rho_I : I \to (0, 1)$ be given by $\rho_I(i_1, \ldots, i_p) = \rho(i_1) \cdot \ldots \cdot \rho(i_p)$. Then if $I$ is tight, by using 2.1(8)(ii), one can check $\sum_{\alpha \in I} \rho_I(\alpha) = 1$, and furthermore
\[ \|S_1, \rho_I\| = \|S, \rho\| \] as follows from 4.4(4) and 3.1(3). Finally, if \( \pi \) is one-one, then 
\[ \|S_1, \rho_I\| = \|S, \rho\| \] iff \( I \) is tight.

5. **Similitudes**

5.1. **Self-Similar Sets.** We continue the notation of §3 and §4.

(1) **Definition.** \( K \) is self-similar (with respect to \( S \)) if

(1) \( K \) is invariant with respect to \( S \), and

(2) \( \mathcal{H}^k(K) > 0 \), \( \mathcal{H}^k(K_i \cap K_j) = 0 \) for \( i \not= j \), where \( k = \dim K \).

Thus (ii) is a kind of “minimal overlap” condition, and rules out the examples in 3.3(1) with \( \frac{1}{2} < r < 1 \). However, it is still rather weak. For example, if \( S = \{ S_1, S_2 \}, \) 
\[ S_1 : \mathbb{R}^2 \to \mathbb{R}^2, \quad S_1(re^{i\theta}) = (\sqrt{2})^{-1}re^{i(\theta - \pi/4)} \] and \( S_2(1+re^{i\theta}) = 1+(\sqrt{2})^{-1}re^{i(\theta + \pi/4)} \), then \( |S| \) is a continuous image of \([0, 1]\) with a dense set of self-intersections, see [OP, D=2, L=0] or [LP, Figure 3]. In [LP, page 374] it is shown \( \mathcal{H}^2(|S|) = \frac{1}{4} \), and so \( \dim |S| = 2 \) and \( |S| \) is self-similar in the above sense by (4)(ii).

A more useful notion of self-similarity may be a condition analogous to the open set condition below, which we will see implies \(|S|\) is self-similar, but allows us to “separate out” the components \(|S_i|\).

(2) **Convention.** For the rest of this section we restrict to the case \( S \) is a family of similitudes. \((X, d)\) will be \( \mathbb{R}^n \) with the Euclidean metric, although other conditions suffice. \( \mathcal{H}^k \) is Hausdorff \( k \)-dimensional measure. \( \text{Lip} (S_i) = r_i \).

Let \( \gamma(t) = \sum_{i=1}^{N} r_i^t \). Then \( \gamma(0) = N \) and \( \gamma(t) \downarrow 0 \) as \( t \to \infty \), and hence there is a unique \( D \) such that \( \sum_{i=1}^{N} r_i^D = 1 \).

(3) **Definition.** If \( \sum r_i^D = 1 \), \( D \) is called the similarity dimension of \( S \).

We will see in 5.3(1) that \( D \) often equals the Hausdorff dimension of \(|S|\).

For the rest of this section \( \rho = \{ \rho_1, \ldots, \rho_N \} \) where \( \rho_i = r_i^D \), and so \( \rho \) is determined by \( S \). We write \( ||S|| \) for \( ||S, \rho|| \), and often write \( K \) for \(|S|\) and \( \mu \) for \(||S||\).

Note that \( \sum_{i_1, \ldots, i_p} r_{i_1}^D \cdots r_{i_p}^D = \left( \sum_{i=1}^{N} r_i^D \right)^p = 1 \), a fact we use frequently.

Finally we take \( r_1 \leq r_2 \leq \cdots \leq r_N \), so that \( r_1 = \min \{ r_i : 1 \leq i \leq N \}, \) 
\( r_N = \max \{ r_i : 1 \leq i \leq N \} \).

(4) **Proposition.** Let \( K = |S|, \) \( \dim K = k \). Then

(i) \( \mathcal{H}^D(K) < \infty \) and so \( k \leq D \) (this is true for arbitrary contractions \( S_i \)).

(ii) \( 0 < \mathcal{H}^k(K) < \infty \) implies \( K \) is self-similar iff \( k = D \).

Proof. (i) \( K = \bigcup_{i_1, \ldots, i_p} K_{i_1} \cdots i_p \) and \( \sum_{i_1, \ldots, i_p} (\text{diam } K_{i_1} \cdots i_p)^D = \sum_{i_1, \ldots, i_p} r_{i_1}^D \cdots r_{i_p}^D (\text{diam } K)^D = (\text{diam } K)^D. \) Since \( \text{diam } K_{i_1} \cdots i_p \leq r_N^D \text{diam } K \to 0 \) as \( p \to \infty \), we are done.

(ii) Suppose \( 0 < \mathcal{H}^k(K) < \infty \) and \( K \) is self-similar, so that \( \mathcal{H}^k(K_i \cap K_j) = 0 \) if \( i \not= j \). Then \( \mathcal{H}^k(K) = \sum_{i=1}^{N} \mathcal{H}^k(K_i) = \sum_{i=1}^{N} r_i^k \mathcal{H}^k(K), \) hence \( \sum r_i^k = 1 \), hence \( D = k. \)

Conversely, suppose \( 0 < \mathcal{H}^D(K) < \infty \). Then \( \mathcal{H}^D(K) \leq \sum_{i=1}^{N} \mathcal{H}^D(K_i) = \sum_{i=1}^{N} r_i^D \mathcal{H}^D(K). \) Since \( \sum_{i=1}^{N} r_i^D = 1, \mathcal{H}^D(K) = \sum_{i=1}^{N} \mathcal{H}^D(K_i) \) and so it is standard measure theory that \( \mathcal{H}^D(K_i \cap K_j) = 0 \) if \( i \not= j. \) \( \square \)

5.2. **Open Set Condition.** Recall convention 5.1(2).

(1) **Definition.** \( S \) satisfies the open set condition if there exists a non-empty open set \( O \) such that

(i) \( \bigcup_{i=1}^{N} S_i O \subset O, \)

(ii) \( S_i O \cap S_j O = \emptyset \) if \( i \not= j. \)
Proof. (a) Suppose we already have a non-empty closed set \( C \) satisfying (i) and (ii) of (1) with \( O \) replaced by \( C \). Let \( d = \min_{i \neq j} d(S(C), S(C)) \), and select \( \varepsilon \) so \( r_1, \varepsilon < a/2 \) for \( i = 1, \ldots, N \). Then \( S \) satisfies the open set condition with \( O = \bigcup_{x \in C} B(x, \varepsilon) \). To see this observe that

\[
S_i O = \bigcup_{x \in C} S_i B(x, \varepsilon) = \bigcup_{x \in C} B(S(x, r_i \varepsilon)) = \bigcup_{y \in S, C} B(y, r_i \varepsilon).
\]

Hence \( S_i O \cap S_j O = \emptyset \) if \( i \neq j \) and furthermore \( S_i O \subseteq O \). This situation applies to 3.3(1), \( 0 < r < \frac{1}{5} \), with the non-empty closed set \( [0, 1] \).

(b) Suppose there is a closed set \( C \) with non-empty interior such that

\[
\begin{align*}
(1) & \quad S_i C \subseteq C \text{ if } i = 1, \ldots, N, \\
(2) & \quad (S_i C)^\circ \cap (S_j C)^\circ = \emptyset \text{ if } i \neq j.
\end{align*}
\]

Then the open set condition holds with \( O = C^\circ \). This situation applies in 3.3(1), (2) with \( C \) the closed convex hull of \( |S| \), i.e. \( C = [0, 1] \) for 3.3(1) and \( C \) is the triangle \((a_1, a_2, a_3)\) for 3.3(2).

(3) Elementary consequences. Suppose \( S \) satisfies the open set condition with \( O \). Note that \( S_{i_1 \cdots i_p} \) commutes with the topological operators \(-,\partial,\partial^\circ\). In particular \((O_{i_1 \cdots i_p})^{-1} = (O^{-1})_{i_1 \cdots i_p}\), and so we can write \( O_{i_1 \cdots i_p} \) unambiguously.

Then

\[
\begin{align*}
(i) & \quad O \supseteq O_{i_1} \supseteq O_{i_1 i_2} \supseteq \cdots \supseteq O_{i_1 \cdots i_p} \supseteq \cdots; \\
(ii) & \quad K_{i_1 \cdots i_p} \subseteq O_{i_1 \cdots i_p}; \\
(iii) & \quad K_{j_1 \cdots j_p} \cap O_{i_1 \cdots i_p} = \emptyset \text{ if } (j_1, \ldots, j_p) \neq (i_1, \ldots, i_p); \\
(iv) & \quad \text{if } I \text{ is tight (2.1(7)), then the } O_{i_\alpha} \text{, } \alpha \in I, \text{ are mutually disjoint.}
\end{align*}
\]

Thus (ii) and (iii) say that \( O_{i_1 \cdots i_p} \) “isolates” \( K_{i_1 \cdots i_p} \) from the \( K_{j_1 \cdots j_p} \) for \((j_1, \ldots, j_p) \neq (i_1, \ldots, i_p)\).

Proof. (i) and (ii) follow immediately from 3.1(8).

For (iii), suppose \((j_1, \ldots, j_p) \neq (i_1, \ldots, i_p)\). But \( K_{j_1 \cdots j_p} \subseteq O_{j_1 \cdots j_p} \), and \( O_{j_1 \cdots j_p} \cap O_{i_1 \cdots i_p} = \emptyset \) since \( O_{j_1 \cdots j_p} \cap O_{i_1 \cdots i_p} = \emptyset \).

For (iv), suppose \( I \) is tight, \( \alpha, \beta \in I, \) and \( \alpha \neq \beta \). Let \( p \) be the greatest integer (perhaps 0) for which there is a sequence \((i_1, \ldots, i_p)\) with \((i_1, \ldots, i_p) \prec \alpha \) and \((i_1, \ldots, i_p) \prec \beta \). Since \( I \) is tight there exist \( i_{p+1} \neq j_{p+1} \) such that \((i_1, \ldots, i_{p+1}) \prec \alpha \), \((i_1, \ldots, i_{p+1}) \prec \beta \). But then \( O_{\alpha} \subset O_{i_1 \cdots i_{p+1}} \), \( O_{\beta} \subset O_{i_1 \cdots i_{p+1}} \) by (1), and so

\[
O_{\alpha} \cap O_{\beta} \subset S_{i_1 \cdots i_p} (O_{i_{p+1}} \cap O_{j_{p+1}}) = \emptyset.
\]

5.3. Existence of Self Similar Sets.

(1) Theorem. Suppose \( S \) satisfies the open set condition. Then

\[
\begin{align*}
(i) & \quad \text{there exist } \lambda_1, \lambda_2 \text{ such that } 0 < \lambda_1 \leq \theta^O(K, k) \leq \theta^O(K, k) \leq \lambda_2 < \infty \text{ for all } k \in K, \\
(ii) & \quad 0 < H^D(K) < \infty \text{ and so } K \text{ is self-similar by 5.1(4)(ii). In particular dim } K = D, \\
(iii) & \quad ||S|| = [H^D(K)]^{-1} H^D[K].
\end{align*}
\]

Proof.

(a) Lemma. Suppose \( 0 < c_1 < c_2 < \infty \) and \( 0 < \rho < \infty \). Let \( \{U_i\} \) be a family of disjoint open sets. Suppose each \( U_i \) contains a ball of radius \( \rho c_1 \) and is contained in a ball of radius \( \rho c_2 \). Then at most \((1 + 2c_2) c_1^{-n}\) of the \( U_i \) meet \( B(0, \rho) \).
For suppose $U_1, \ldots, U_k$ meet $B(0, \rho)$. Then each of $U_1, \ldots, U_k$ is a subset of $B(0, (1 + 2c_2)\rho)$. Summing the volumes of the $k$ corresponding disjoint spheres of radius $\rho c_1$, we see that
\[ k \alpha_n \rho^n c_1^n \leq \alpha_n (1 + 2c_2)^n \rho^n, \]
and hence $k \leq (1 + 2c_2)^n c_1^{-n}$.

(b) For the rest of the proof let $O$ be the open set asserted to exist by the open set condition.

Let $\mu = \|S\|$. We will first prove that there exist constants $\kappa_1, \kappa_2$ such that
\[ 0 < \kappa_1 \leq \theta^s(\mu, k) \leq \theta^{sD}(\mu, k) \leq \kappa_2 < \infty \]
for all $k \in K$.

First note that
\[ \mu(K_{i_1} \ldots i_p) \geq \mu_{i_1 \ldots i_p}(K_{i_1} \ldots i_p) = r_{i_1}^D \cdots r_{i_p}^D \mu(S_{i_1}^{-1} K_{i_1} \ldots i_p) = r_{i_1}^D \cdots r_{i_p}^D \mu(K) = r_{i_1}^D \cdots r_{i_p}^D. \]
Let $k = k_{i_1 \ldots i_p} \ldots$ and consider $B(k, \rho)$. Choose the least $p$ such that $K_{i_1} \ldots i_p \subset B(k, \rho)$. Then $r_{i_1} \cdots r_{i_p} (\text{diam } K) \geq \rho r_1$ (recalling $r_1 \leq \cdots \leq r_N$). Hence
\[ \frac{\mu(B(k, \rho))}{\alpha_D \rho^D} \geq \frac{\mu(K_{i_1} \ldots i_p)}{\alpha_D \rho^D} \geq \frac{r_{i_1}^D \cdots r_{i_p}^D}{\alpha_D \rho^D (\text{diam } K)^{-D}}. \]
Hence $\theta^D(\mu, k) \geq r_{i_1}^D \alpha_D^{-1} (\text{diam } K)^{-D}$ for $k \in K$.

We now show that $\theta^{sD}(\mu, k)$ is uniformly bounded away from $\infty$ for $k \in K$.

Suppose $O$ contains a ball of radius $c_1$ and is contained in a ball of radius $c_2$. For each sequence $j_1 \ldots j_q \ldots \in C(N)$ select the least $q$ such that $r_1 \rho \leq r_{j_1} \cdots r_{j_q} \leq \rho$. Let $I$ be the set of $(j_1, \ldots, j_q)$ thus selected, and notice that $I$ is tight (2.1(7)). From 5.2(3) it follows $\{O_{j_1, \ldots, j_q} : (j_1, \ldots, j_q) \in I\}$ is a collection of disjoint open sets. Moreover, each such $O_{j_1, \ldots, j_q}$ contains a ball of radius $r_{j_1} \cdots r_{j_q} c_1$ and hence of radius $r_1 c_1$ and is contained in a ball of radius $r_{j_1} \cdots r_{j_q} c_2$ and hence of radius $\rho c_2$. It follows from (a) that at most $(1 + 2c_2)^n (r_1 c_2)^{-n}$ of the $O_{j_1, \ldots, j_q}$, $(j_1, \ldots, j_q) \in I$, meet $B(k, \rho)$. Hence at most $(1 + 2c_2)^n (r_1 c_2)^{-n}$ of the $K_{j_1, \ldots, j_q}$, $(j_1, \ldots, j_q) \in I$, meet $B(k, \rho)$.

Now $\text{spt}(\mu_{j_1, \ldots, j_q}) = K_{j_1, \ldots, j_q}$ by 4.4(4)(ii). By 4.5
\[ \mu = \sum_{(j_1, \ldots, j_q) \in I} \mu_{j_1, \ldots, j_q}. \]
Finally $M(\mu_{j_1, \ldots, j_q}) = r_{j_1}^D \cdots r_{j_q}^D \leq \rho^D$ for $(j_1, \ldots, j_q) \in I$.

Hence
\[ \frac{\mu(B(k, \rho))}{\alpha_D \rho^D} \leq \frac{(1 + 2c_2)^n}{r_{j_1}^D \cdots r_{j_q}^D \alpha_D \rho^D} = \frac{(1 + 2c_2)^n}{\alpha_D r_{j_1}^D \cdots r_{j_q}^D}. \]
It follows $\theta^{sD}(\mu, k) \leq (1 + 2c_2)^n (\alpha_D r_{j_1}^D \cdots r_{j_q}^D)^{-1}$.

(c) (ii) now follows from 2.6(3).

(d) Since $K$ is self-similar, $H^D(K_i \cap K_j) = 0$ if $i \neq j$, and so
\[ H^D[K] = \sum_{i=1}^{N} H^D[K_i] = \sum_{i=1}^{N} r_{i}^D S_{i \#}(H^D[K]) \]
by 2.6(2).

Letting $\tau = [H^D(K)]^{-1} H^D[K]$, it follows that $\tau = \sum_{i=1}^{N} r_{i}^D S_{i \#}(\tau)$, and that $M(\tau) = 1$. By uniqueness, $\tau = \mu$, proving (iii).

(e) From (iii), $\theta^D(K, k) = \theta^D(H^D[K], k) = [H^D(K)]^{-1} \theta^D(\mu, k)$, and similarly for $\theta^{sD}$. (i) now follows from (b). \qed
(2) **Remarks.** Result (i) says that $K$ is rather uniformly spread out in the dimension $k$. But on the other hand, by a result of Marstrand [MJ], the inequality between the upper and lower densities cannot be replaced by an equality if $D$ is non-integral.

Result (ii) is due to Moran [MP].

5.4. **Purely Unrectifiable Sets.** We continue our assumptions that $(X,d)$ is $\mathbb{R}^n$ with the Euclidean metric and that $\mathcal{S}$ is a family of similitudes.

(1) **Theorem.** Suppose $\mathcal{S}$ satisfies the open set condition with both the open set $O$ and the open set $U$, where $O \subset U$. Suppose furthermore that whenever $A$ is an $m$-dimensional affine subspace of $\mathbb{R}^n$ for which $A \cap \tilde{O}_i \neq \emptyset$ and $A \cap \tilde{O}_j \neq \emptyset$ for some $i \neq j$, then $A \cap \left( U \sim \bigcup_{i=1}^N \tilde{O}_i \right) \neq \emptyset$. Then for any $m$-dimensional $C^1$ manifold $M$ in $\mathbb{R}^n$, $\mathcal{H}^m(M \cap |\mathcal{S}|) = 0$.

**Proof.** We proceed in stages.

(a) Let

$$
\mathcal{A} = \{ A : A \text{ is an } m\text{-dimensional affine space with } A \cap \tilde{O}_i \neq \emptyset, A \cap \tilde{O}_j \neq \emptyset \text{ for some } i \neq j \}.
$$

(see Figure 5.1.)

Let $g : \mathcal{A} \to (0, \infty)$ be defined by

$$
g(A) = \sup \left\{ r : B(a, r) \subset U \sim \bigcup_{i=1}^N \tilde{O}_i \text{ for some } a \in A \right\}.
$$

By the hypotheses of the theorem, $0 < g(A) < \infty$. We want to show that:

$g$ is uniformly bounded away from 0.

To do this we define a topology on $\mathcal{A}$ and prove that $\mathcal{A}$ is compact and $g$ is lower semi-continuous (i.e. $g(A_0) > \lambda$ implies $g(A) > \lambda$ for all $A$ sufficiently close to $A_0$).

The required result then follows.

Let $O_m$ be the set of $m$-dimensional subspaces through the origin, give $O_m$ its usual compact topology as a subset of $\mathbb{R}^n^\ast$, and let $\mathbb{R}^n \times O_m$ have the product
topology. The map \( (a, O) \mapsto a + O \) is a map from \( \mathbb{R}^n \times \mathcal{O}_m \) onto the set of \( m \)-dimensional affine spaces in \( \mathbb{R}^n \), we give this set the induced topology. Since \( \mathcal{A} \) is the image of a closed bounded (hence compact) set, it is compact.

Next suppose \( g(A_0) > \lambda, A_0 \in \mathcal{A} \). Select \( a_0 \in A_0 \) and \( \lambda_0 > \lambda \) such that \( B(a_0, \lambda_0) \subset U \sim \bigcup_{i=1}^N O_i \). For all \( A \) sufficiently close to \( A_0 \), \( d(a_0, A) < \lambda_0 - \lambda \). Select \( a \in A \) such that \( d(a_0, a) < \lambda_0 - \lambda \). Then \( B(a, \lambda) \subset B(a_0, \lambda_0) \subset U \sim \bigcup_{i=1}^N O_i \).

Hence \( g(A) > \lambda \). Thus \( g \) is lower semi-continuous. The required result follows.

(b) For each \( \varepsilon > 0 \), let
\[
C_{\varepsilon} = \{ C : C \text{ is an } m\text{-dimensional } C^1 \text{ manifold in } \mathbb{R}^n, \text{ and for some } A \in \mathcal{A}, \text{ there exists a } C^1 \text{ map } f : A \cap U \rightarrow C \text{ such that } \]
\[
\begin{align*}
& (i) \ f \text{ is one-one}, \\
& (ii) \ \text{Lip } f \leq 1 + \varepsilon, \ \text{Lip } f^{-1} \leq 1 + \varepsilon, \\
& (iii) \ d(f(x), x) < \varepsilon \text{ for all } x \in A \cap U \}. 
\end{align*}
\]

We will show:
\[
\text{there exist } \varepsilon > 0, \delta > 0 \text{ such that } \]
\[
\mathcal{H}^m \left( C \cap \left( U \sim \bigcup_{i=1}^N O_i \right) \right) > \delta \text{ for all } C \in C_{\varepsilon}.
\]

Suppose by (a) that \( g(A) > \lambda > 0 \) for all \( A \in \mathcal{A} \). Fix \( \varepsilon \) so that \( 0 < \varepsilon < \lambda/3 \).

Suppose \( C \in C_{\varepsilon} \) with \( f \) and \( A \) as in the definition of \( C_{\varepsilon} \). Select \( a \in A \) such that \( B(a, \lambda) \subset U \sim \bigcup_{i=1}^N O_i \).

Since \( d(f(a), a) < \varepsilon \) it follows \( B(f(a), \lambda - \varepsilon) \subset B(a, \lambda) \) and hence \( B(f(a), \lambda - \varepsilon) \subset U \sim \bigcup_{i=1}^N O_i \). But one can check that \( f(A \cap B(a, \lambda - 3\varepsilon)) \subset C \cap B(f(a), \lambda - \varepsilon) \). It follows that
\[
\mathcal{H}^m \left( C \cap \left( U \sim \bigcup_{i=1}^N O_i \right) \right) \geq \int_{A \cap B(a, \lambda - 3\varepsilon)} J(f) \, d\mathcal{H}^m \geq \alpha_m (\lambda - 3\varepsilon)^m (1 + \varepsilon)^{-m},
\]
where \( J(f) \) is the Jacobian of \( f \). This gives the required result.

(c) Now assume that the hypotheses of the theorem hold and that \( \mathcal{H}^m(M \cap K) \neq 0 \) for some \( C^1 \) manifold \( M \), where \( K = [S] \). We will deduce a contradiction.

First note that \( \theta^m(M \cap K, k) = 1 \) for some \( k \in M \cap K \) [FH, 3.2.19] since \( M \cap K \) is \( (\mathcal{H}^m, m) \)-rectifiable. Alternatively the corresponding result for \( \theta^m \) in \( \mathbb{R}^m \) [MM, page 184] can readily be lifted back to the manifold \( M \) by means of the area formula
\[
\mathcal{H}^m(f(A)) = \int_A J(f) d\mathcal{L}^m \text{ for } C^1 \text{ diffeomorphisms } f : A \rightarrow \mathbb{R}^m, A \subset \mathbb{R}^m.
\]

In the following we will need to be a little careful, since due to “overlap” it is possible that for fixed \( p \), an arbitrary member of \( K \) may belong to more than one \( K_{i_1 \ldots i_p} \).

Since \( k \) is a point of non-zero \( m \)-dimensional density for \( M \cap K \), it follows that there is a sequence \( k_j \rightarrow k \) as \( j \rightarrow \infty, k \neq k_j \in M \cap K \). By passing to a subsequence we may suppose all \( k_j \in K_{i_j} \) for some \( i_j \), which we fix. By passing to a subsequence again we may suppose all \( k_j \in K_{i_1 i_2} \) for some \( i_2 \) which we also fix. Repeating this argument and then diagonalising, we extract a subsequence \( k_j \rightarrow k \) and a sequence \( i_1 \ldots i_j \ldots \), such that \( k_j \in K_{i_1 \ldots i_j} \) for all \( j \). Moreover \( k = k_{i_1 \ldots i_j} \ldots \)

For each \( j \) let \( p(j) \) be the least integer \( p \) such that \( k_j \in K_{i_{1 \ldots i_p}}, k_j \notin K_{i_{1 \ldots i_p} i_{p+1}}, \) and notice \( p(j) \geq j, \) so \( p(j) \rightarrow \infty \) as \( j \rightarrow \infty \).

Now \( \theta^m(M \cap K, k) = 1 \), and \( \theta^m(M, k) = 1 \) since \( M \) is a \( C^1 \) manifold, hence \( \theta^m(M \sim K, k) = 0 \). Select \( R \geq \text{diam } U \), so that \( \text{diam } U_{i_{1 \ldots i_p(j)}} < Rr_{i_1}, \ldots, r_{i_{p(j)}} \).
We have
\[
\lim_{j \to \infty} \frac{\mathcal{H}^m((M \sim K) \cap B(k, Rr_{1j} \cdots r_{ip(j)}))}{\alpha_m(Rr_{1j} \cdots r_{ip(j)})^m} = 0.
\]

To simplify notation we write \( p \) for \( p(j) \). See Figure 5.2.

Figure 5.2

Now
\[
(M \sim K) \cap B(k, Rr_{1j} \cdots r_{ip}) \supset (M \sim K) \cap U_{1j} \cdots p = M \cap (U_{1j} \cdots r_{ip}) \sim K
\]
by 5.2(3)(iii)
\[
M \cap (U_{1j} \cdots r_{ip}) \sim \bigcup_{\alpha=1}^N \bar{O}_{1j} \cdots r_{ip \alpha}) \text{ by 5.2(3)(ii)}.
\]

Hence
\[
\frac{\mathcal{H}^m((M \sim K) \cap B(k, Rr_{1j} \cdots r_{ip(j)}))}{\alpha_m(Rr_{1j} \cdots r_{ip(j)})^m} \geq \frac{\mathcal{H}^m(M \cap (U_{1j} \cdots r_{ip(j)} \sim \bigcup_{\alpha=1}^N \bar{O}_{1j} \cdots r_{ip \alpha}))}{\alpha_m(Rr_{1j} \cdots r_{ip(j)})^m} = \mathcal{H}^m(f_j(M) \cap (U \sim \bigcup_{\alpha=1}^N \bar{O}_{\alpha})),
\]

where \( f_j = S_{ip(j)}^{-1} \circ \cdots \circ S_{ip(j)}^{-1} \) is an “explosion” map. Here we are using the fact that
\[
f_j(U_{1j} \cdots p(j)) = U, \quad f_j(O_{1j} \cdots p(j) \alpha) = \bar{O}_{\alpha}, \quad \mathcal{H}^m(f_j(A)) = r_{ip(j)}^{-m} \cdots r_{1j}^{-m} \mathcal{H}^m(A)
\]
for arbitrary \( A \).

But for sufficiently large \( j \) we will show that \( f_j(M) \in C_\varepsilon \). From (b) this shows the expression in (b) is bounded away from 0 for all sufficiently large \( j \), contradicting (a). Thus our original assumption that \( \mathcal{H}^m(M \cap K) \neq 0 \) is false.

To see that \( f_j(M) \in C_\varepsilon \) for all sufficiently large \( j \) we first observe that in analogy with the definition of \( C_\varepsilon \), we have for all sufficiently large \( j \) a \( C^1 \) map \( g_j : f_j(T) \cap V \to f_j(M) \) where

(i) \( V \) is a fixed open neighbourhood of \( \bar{U} \),
(ii) \( T \) is the tangent plane to \( M \) at \( k \),
(iii) \( g_j \) is one-one,
(iv) \( \operatorname{Lip} g_j \leq 1 + \varepsilon, \operatorname{Lip} g_j^{-1} \leq 1 + \varepsilon \).
(v) \( d(g_j(x), x) < \varepsilon/2 \) for all \( x \in f_j(T) \cap V \).

See for example [FH, 3.1.23]. But we do not know if \( f_jM \in \mathcal{A} \). However, we can select an affine space \( A_j \) through \( k \) and \( k_j \) such that for all sufficiently large \( j \), \( f_j(A_j) \cap U \) and \( f_j(T) \cap V \) are arbitrarily close in the topology on affine spaces introduced in (a). In particular there will exist \( \psi_j : f_j(A_j) \cap U \rightarrow f_j(T) \cap V \) such that \( \text{Lip } \psi_j \leq 1 \) and \( d(\psi_j(x), x) < \varepsilon/2 \) for all \( x \in f_j(A_j) \cap U \).

Notice that \( f_j(A_j) \in \mathcal{A} \), since \( f_j(k) \in f_j(A_j) \cap K_{iM_{(j)j+1}} \subset A \cap \hat{O}_k \) and \( f_j(k_j) \in f_j(A_j) \cap K_{\alpha} \subset A \cap \hat{O}_k \) for some \( \alpha \neq iM_{(j)j+1} \).

Finally \( f_j(M) \in \mathcal{C}_\varepsilon \), where in the definition of \( \mathcal{C}_\varepsilon \), \( A \) is replaced by \( f_j(A_j) \), \( f \) is replaced by \( g_j \circ \psi_j \), and \( C \) is replaced by \( f_j(M) \). \( \square \)

(2) **Remark.** In the terminology of [FH, 3.2.14], \( K \) is purely \((H^m, m)\) unrectifiable. In this respect the interest of the present theorem lies in the fact that it establishes pure \((H^m, m)\) unrectifiability for sets \( K \) such that \( H^m(K) = \infty \) (provided \( m < D \)). Unlike the examples in [FH, 3.3.19, 3.3.20] one cannot argue by using the structure theorems for sets having finite \( H^m \) measure.

(3) **Example.**

(a) If \( K = U \cup_k \) with the \( K_i \) disjoint, then the hypotheses of the theorem are easily seen to be satisfied if \( m = 1 \), where \( O \) is as in 5.2(2)(a), and \( U = O \). For \( A \cap (U \sim \bigcup_{i=1}^j O_i) = \emptyset \) if \( A \subset (U \sim \bigcup_{i=1}^j O_i)^c = U^c \cup \bigcup_{i=1}^j O_i \). But this latter cannot be true if \( A \in \mathcal{A} \), since then \( A \) can be split into two disjoint non-empty components \( A \cap U^c \) and \( A \cap \bigcup_{i=1}^j O_i \).

(b) From Example 3.3(2) let \( O \) be the interior of the triangle \((a_1, a_5, a_3)\). Let \( U \supset O \) be a slightly larger open set also satisfying the open set condition and such that \( \partial U \cap \partial O = \{a_1, a_3\} \). Suppose \( A \in \mathcal{A} \) where \( m = 1 \). If \( A \cap (U \sim \bigcup_{i=1}^j O_i) = \emptyset \) then it is straightforward to show \( \{a_1, a_3\} \subset A \). But then \( \{a_2, a_4\} \subset A \), which contradicts \( A \cap (U \sim \bigcup_{i=1}^j O_i) = \emptyset \) since \( \{a_2, a_4\} \subset U \sim \bigcup_{i=1}^j O_i \).

(4) **Remark.** Let us strengthen the hypotheses in (1) by taking \( A \) to be a one-dimensional affine subspace. Then Mattila [MaP] has shown the existence of an \( \varepsilon > 0 \), depending only on \( S \), such that for any \( m \)-dimensional \( C^1 \) manifold \( M \), \( \dim (M \cap |S|) \leq m - \varepsilon \).

In the same paper, Mattila also shows that under the hypotheses of 5.3(1), if \( m \geq D \), then there are only two possibilities; either \( K \) lies in an \( m \)-dimensional affine subspace or \( H^D(|S| \cap M) = 0 \) for every \( m \)-dimensional \( C^1 \) manifold \( M \).

5.5. **Parameter Space.** The orthogonal group \( O(n) \) of orthonormal transformations of \( \mathbb{R}^n \) is an \( (n(n-1))/2 \) dimensional manifold [FH, 3.2.28(1)], and hence the set of similitudes \( S = \{a, r, O\} \) in \( \mathbb{R}^n \) corresponds to an \( n(n-1)/2 + (n+1) = (n^2 + n + 2)/2 \) dimensional manifold. Thus every invariant set generated by some \( S = \{S_1, \ldots, S_N\} \) of similitudes in \( \mathbb{R}^n \) corresponds to a point in an \( N(n^2 + n + 2)/2 \) dimensional manifold, which we call the parameter space. Oppenheimer [OP] has made a systematic computer analysis of a part of \( N = n = 2 \).

6. **Integral Flat Chains**

In this section we will see how integral flat chains, which will not normally be rectifiable, arise naturally in the context of self-similarity. In particular, the Koch curve of 3.3(2) supports a \( 1 \)-dimensional integral flat chain in a natural way, and \( |S| \) in Example 3.3(3) supports a 2-dimensional integral flat chain provides \( D < 3 \) (with \( D \) defined in 6.2(2)).

We make the convention that all currents we consider are integral flat chains, or chains for short.

We need to introduce a new metric, but first we need a lemma on the \( F \)-metric.
6.1. The $\mathcal{F}$-metric.

(1) **Lemma.** Suppose $1 \leq m \leq n-1$, $T$ is a (not necessarily rectifiable) $m$-cycle, and $\gamma = \gamma(m,n)$ is the isoperimetric constant of $2.7(5)$.

(i) If $\mathcal{F}(T) < \gamma^{-m}$, $T = \partial A + R$, and $\mathcal{F}(T) = \mathcal{M}(A) + \mathcal{M}(R)$, then $R = 0$.

(ii) If $\mathcal{F}(T) \leq \gamma^{-m}$, then $T = \partial A$ for some $A$ such that $\mathcal{M}(A) = \mathcal{F}(T)$.

**Proof.** (i) We first remark that any $T$ can be written as $T = \partial A + R$ with $\mathcal{F}(T) = \mathcal{M}(A) + \mathcal{M}(R)$, as noted in $2.7(5)$.

Assume the hypotheses of (i). If $R = 0$ we are done, so suppose $R \neq 0$. Since $\partial R = \partial T = 0$, and $\mathcal{M}(R) \leq \mathcal{F}(T) < \gamma^{-m}$, there is an $m$-cycle $D$ such that $R = \partial D$ and $\mathcal{M}(D) \leq \gamma \mathcal{M}(R)^{m+1/m}$ by $2.7(5)$. Hence $\mathcal{M}(D) < \mathcal{M}(R)$, since $[\mathcal{M}(R)]^{1/m} \leq [\mathcal{F}(T)]^{1/m} < \gamma^{-1}$. But then $T = \partial(A + D) = \mathcal{M}(A) + \mathcal{M}(D) < \mathcal{M}(A) + \mathcal{M}(R) = \mathcal{F}(T)$, a contradiction.

(ii) Suppose $\mathcal{F}(T) \leq \gamma^{-m}$ and let $T = \partial A + R$ with $\mathcal{F}(T) = \mathcal{M}(A) + \mathcal{M}(R)$. Then the same argument as for (i) shows that $R = \partial D$ with $\mathcal{M}(D) < \mathcal{M}(R)$. Hence $T = \partial(A + D)$ and $\mathcal{M}(A + D) \leq \mathcal{M}(A) + \mathcal{M}(D) \leq \mathcal{M}(A) + \mathcal{M}(R) = \mathcal{F}(T)$. Thus $\mathcal{M}(A + D) = \mathcal{F}(T)$, and so we are done.

(2) We see the necessity of the condition $\mathcal{F}(T) \leq \gamma^{-m}$ in the following example. Let $T_\gamma$ be a 1-cycle supported on $\{x : |x| = \gamma\} \subset \mathbb{R}^2$ with $\mathcal{M}(T_\gamma) = 2\pi \gamma$. Let $A_\gamma$ be the rectifiable 2-current supported on $\{x : |x| \leq \gamma\} \subset \mathbb{R}^2$ such that $\partial A_\gamma = T$ and $\mathcal{M}(A_\gamma) = \pi \gamma^2$. Using the fact $\gamma(1,2) = 4\pi$ [FH 4.5.14] and the constancy theorem [FH 4.1.7], one can show that if $r \leq 2$, then $\pi r^2 = \mathcal{M}(A_r) = \mathcal{F}(T_r)$. If $r > 2$, then again by [FH 4.1.7] $\partial C = T_r$ implies $C = A_r$ and so $\mathcal{M}(C) = \pi r^2$ (unless we allow $C$ to have non-bounded support, in which case $\mathcal{M}(C) = \infty$). But $\pi r^2 > 2\pi r = \mathcal{M}(T_r) \geq \mathcal{F}(T_r)$.

6.2. The $\mathcal{C}$-metric.

(1) **Definition.** Let $B$ be an $(m-1)$-boundary, $m \geq 1$. Then $\mathcal{C}_B$ is the set of $m$-chains given by

$$\mathcal{C}_B = \{R \in \mathcal{F}_m : \partial R = B\}.$$

(2) **Definition.** For $R, S \in \mathcal{C}_B$ let

$$\mathcal{C}(R, S) = \inf \{\mathcal{M}(A) : A \in \mathcal{R}_{m+1}, R - S = \partial A\}.$$

We now see that $\mathcal{C}$ is a complete metric on $\mathcal{C}_B$ with the useful transformation result $(3)(i)$.

(3) **Lemma.** Let $B$ be an $(m-1)$-boundary, $m \geq 1$.

(i) If $f : \mathbb{R}^n \to \mathbb{R}^m$ is a proper Lipschitz map and $\text{Lip } f = r$, then $\mathcal{C}(f \# R, f \# S) \leq r^{m+1} \mathcal{C}(R, S)$.

(ii) $\mathcal{F}(R, S) \leq \mathcal{C}(R, S)$, and $\mathcal{F}(R, S) = \mathcal{C}(R, S)$ if $\mathcal{F}(R, S) \leq \gamma^{-m}$.

(iii) $\mathcal{C}$ is a complete metric on $\mathcal{C}_B$. The $\mathcal{C}$- and $\mathcal{F}$-topologies agree on $\mathcal{C}_B$.

(iv) The infimum in the definition of $\mathcal{C}(R, S)$ is realised for some $A \in \mathcal{R}_{m+1}$.

**Proof.** (i) is immediate from $2.7(6)(c)$.

The first assertion in (ii) is immediate, and the second follows from $6.1(1)(i)$.

To see that $\mathcal{C}(R, S) < \infty$ let $\mathcal{F}(R, S) = \lambda < \infty$ and by $2.7(6)(c)$ choose $f = \mu$, such that $\mathcal{F}(\mu_r \# R, \mu_r \# S) \leq \gamma^{-m}$ ($\mu_r$ is defined in $2.3$). Then $\mathcal{C}(\mu_r \# R, \mu_r \# S) \leq \gamma^{-m}$ and so by (i) $\mathcal{C}(R, S) \leq r^{-(m+1)} \gamma^{-m}$. The other properties of a metric are easily verified, noting in particular that if $\mathcal{C}(R, S) = 0$ then $\mathcal{F}(R, S) = 0$ and so $R = S$.

The $\mathcal{C}$- and $\mathcal{F}$-topologies clearly agree on $\mathcal{C}_B$. Since a sequence is $\mathcal{C}_B$-Cauchy if it is $\mathcal{F}$-Cauchy, $\mathcal{C}_B$ is closed in $\mathcal{F}_m$ in the $\mathcal{F}$-metric, and $\mathcal{F}$ is a complete metric on $\mathcal{F}_m$, it follows $\mathcal{C}$ is a complete metric on $\mathcal{C}_B$. 


To prove (iv) suppose \( C(R, S) = \lambda \) and let \( T_j \in \mathcal{R}_{m+1}, \partial T_j = R - S, M(T_j) = \lambda \). Let \( B(0, r) \) be some ball large enough to include \( \text{spt} \ (R - S) \), and let \( f : \mathbb{R}^n \to \mathbb{R} \) be a retraction map with Lipschitz constant 1. Let \( T'_j = f \# T_j \). Then \( \text{spt} \ T'_j \subset B(0, r) \) and \( M(T'_j) \leq M(T_j) \) by 2.7(6). We can apply the compactness theorem of 2.7(5) to \( T'_j \to T'_1 \) and extract a convergent subsequence with limit \( A - T'_1 \), say. From 2.7(6) it follows \( M(A) \leq \lambda \) and hence \( M(A) = \lambda \). Furthermore \( \partial A = R - S \), and so \( A \) is the required current.

\[ \square \]

6.3. Invariant Chains. Suppose \( S = \{S_1, \ldots, S_N\} \) are proper contraction maps on \( \mathbb{R}^n \), not necessarily similitudes.

1. **Definition.** For any \( k \)-chain \( T \), we let \( S(T) = \sum_{i=1}^N S_i(T) \); also \( S^0(T) = T, S^1(T) = S(T), S^{p+1}(T) = S(S^p(T)) \) if \( p \geq 1 \).

2. Suppose now that \( \text{Lip} S_i \leq r_i \), and let \( D \) be specified by \( \sum_{i=1}^N r_i \geq 1 \) as in 5.1(3), but recall that here the \( S_i \) are not necessarily similitudes. Let \( m \) be the unique integer given by \( m \leq D < m + 1 \). Now suppose \( m \geq 1 \), \( B \) is an \((m - 1)\)-boundary, and \( S(B) = B \). As examples consider 3.3(2) with \( m = 1 \) and \( B = [(1, 0)] - [(0, 0)] \), or 3.3(3) with \( m = 2 \) and \( B = N. \) Finally let \( \theta = \sum_{i=1}^N r_i^m \) and note that \( \theta < 1 \).

3. **Theorem.** Under the hypotheses of (2) the following hold

   (i) \( S \) is a contraction map on \( C_B \) in the \( C \)-metric.

   (ii) There is a unique \( m \)-chain \( T \in C_B \) such that \( S(T) = T \).

   (iii) If \( R \in C_B \), then \( S^p(R) \to T \) in the \( F \)-metric (and \( C \)-metric).

   (iv) If \( R \in C_B \) and \( S(R) - R = \partial A \) with \( A \in \mathcal{R}_{m+1} \) (which is always possible by 6.2(3)) then \( A_0 = \sum_{p=0}^\infty S^p(A) \in \mathcal{R}_{m+1} \) with convergence in the \( M \)-norm, and \( T = R + \partial A_0 \).

**Proof.** First note that for any \( D \in F_{m+1} \),

\[
M(S(D)) = M \sum_{i=1}^N S_i(D) \leq \sum_{i=1}^N M(S_i(D)) \leq \sum_{i=1}^N \theta^m(D) \quad \text{(by 2.7(6)(c))} = M(D).
\]

We now show (i). If \( R \in C_B \) then \( S(R) \in C_B \) since \( \partial(S(R)) = S(\partial R) = S(B) = B \). Next suppose by 6.2(3)(iv) that \( R_1 - R_2 = \partial C \) with \( C_B(R_1, R_2) = M(C) \). Then \( S(R_1 - R_2) = S(\partial C) = \partial(S(C)) \), and \( M(S(C)) \leq M(C) \). Hence \( C_B(S(R_1), S(R_2)) \leq C_B(R_1, R_2) \).

(iii) and (i) follow immediately, using 6.2(3).

To establish (iv), suppose \( R \in C_B, S(R) - R = \partial A, A \in \mathcal{R}_{m+1} \). Then \( M(S^p(A)) \leq M(M(S^{p-1}(A))) \) and so \( M(S^p(A)) \leq \theta^p M(A) \). Thus \( A_0 = \sum_{p=0}^\infty S^p(A) \) converges in the \( M \)-norm. Thus \( A_0 \) is a chain of finite mass and hence rectifiable by 2.7(4). Finally

\[
\partial A_0 = \sum_{p=0}^\infty \partial S^p(A) = \sum_{p=0}^\infty S^p(\partial A) = \sum_{p=0}^\infty \partial S^p(S(R) - R) = \lim_{p \to \infty} (S^p(R) - R) = T - R.
\]

\[ \square \]

4. We can often take particularly simple chains for \( R \) and \( A \) in (3)(iv). For example in 3.3(2) we can take \( R = [a_1, a_5] \) and \( A = [a_2, a_3, a_4] \) to be the obvious oriented simplices [FH, 4.1.11].
Again taking the hypotheses of (2), let \( T \in C_B \) be given by (3). Now spt \( T = \text{spt} \ S(T) \subset \bigcup_{i=1}^{N} \text{spt} \ S_i \# T \subset \bigcup_{i=1}^{N} S_i (\text{spt} \ T) = S (\text{spt} \ T) \). Thus \( S^p (\text{spt} \ T) \uparrow \) as \( p \to \infty \), and since the limit in the Hausdorff metric is \( |S| \) it follows \( \text{spt} \ T \subset |S| \).

It is easy to construct examples, where due to “cancellation”, \( \text{spt} \ T \subsetneq |S| \).

**References**


**The Australian National University, Canberra, ACT, 2600, Australia**

*Received March 3, 1980; Revised October 16, 1980*