THE CONTRACTION MAPPING THEOREM, II

KEITH CONRAD

1. Introduction

In part I, we met the contraction mapping theorem and an application of it to solving nonlinear differential equations. Here we will discuss some variations on the contraction mapping theorem and give a second interesting application: the construction of fractals. This will require the use of an abstract metric space, whose "points" are subsets of Euclidean space.

2. Fractals

Fractals like the Cantor set, Sierpinski's triangle, and the Koch curve are usually defined by a recursive modification process starting with a specific figure (for instance, removing middle-thirds starting from the interval $[0,1]$ in the case of the Cantor set). Following an idea of Hutchinson [10], we will describe a different approach to constructing such fractals using the contraction mapping theorem. Our metric space for the contraction will be the set of all nonempty compact subsets of $\mathbf{R}^n$. The "points" of the metric space are nonempty compact subsets of $\mathbf{R}^n$ and the metric giving a distance between compact subsets of $\mathbf{R}^n$ was defined by Hausdorff and is called the Hausdorff metric. Classical fractals arise as fixed points of a contraction on this space of (nonempty) compact subsets of $\mathbf{R}^n$. Much of the technical development we need doesn't require anything special about Euclidean space, so we will work in a general metric space until we have everything in place to return to the application to fractals in $\mathbf{R}^n$ (with $n = 1$ or 2).

Discussions with Monique Ethier were helpful in the preparation of this section.

Let (X, d) be a metric space. Set

$$\mathrm{H}(X) = \text{nonempty closed and bounded subsets of } X.$$

When X is Euclidean space, $\mathrm{H}(X)$ is the set of nonempty compact subsets of X, but in a general metric space the closed and bounded subsets might include some non-compact subsets. We are going to turn $\mathrm{H}(X)$ into a metric space with the following key properties (due to Blaschke): if X is complete then $\mathrm{H}(X)$ is complete, and if X is compact then $\mathrm{H}(X)$ is compact.

To define a metric on $\mathrm{H}(X)$, we use the notion of an expansion of a subset. For a nonempty subset $A \subset X$ and $r \geq 0$, set the *r-expansion* of A to be the points in X within distance at most r of some point of A:

$$E_r(A) = \bigcup_{a \in A} \overline{B}_r(a) = \{x \in X : d(x, a) \leq r \text{ for some } a \in A\}.$$

In particular, if $A = \{a\}$ is a one-element set then the r-expansion $E_r(\{a\})$ is the closed ball $\overline{B}_r(a)$. For any A, $E_0(A) = A$.

Example 2.1. Take $X = \mathbf{R}^2$ with its usual metric. For a point P in $\mathbf{R}^2$, $E_r(\{P\}) = \overline{B}_r(P)$ is a closed ball centered at P. For a circle C in $\mathbf{R}^2$, $E_r(C)$ is an annulus of width $2r$ with C as its middle circle when r is less than the radius of C. When r equals or exceeds the radius of C, $E_r(C)$ is a closed ball. For a line segment L in $\mathbf{R}^2$, $E_r(L)$ is a (flat) cigar-shaped region with L lying along its middle. If P is a point of $\mathbf{R}^2$ not on L then $\{P\} \subset E_r(L)$ as long as r is at least as large as the shortest distance between P and a point of L, but for the minimal choice of r we will not have $L \subset E_r(\{P\})$. Draw a picture! So the relations $\{P\} \subset E_r(L)$ and $L \subset E_r(\{P\})$ are not the same in general.

We define the distance between two subsets of X, roughly speaking, to be the smallest r such that each set is contained in the r-expansion of the other set. For a precise definition, we use infimums in place of minimums and restrict attention to (nonempty) closed and bounded subsets of X.

Definition 2.2. For $A, B \in \mathrm{H}(X)$, their *Hausdorff distance* is

$$d_{\mathrm{H}}(A, B) = \inf\{r \geq 0 : A \subset E_r(B) \text{ and } B \subset E_r(A)\}.$$

Let's check the definition makes sense. That is, for bounded subsets A and B of X we want to show $A \subset E_r(B)$ and $B \subset E_r(A)$ for some r, so there really are r's to be taking infimums over. Since A and B are bounded, $A \subset \overline{B}_R(x)$ and $B \subset \overline{B}_{R'}(x')$ for some x and x' in X and $R, R' > 0$. By the triangle inequality, for any $a \in A$ and $b \in B$ we have

$$d(a, b) \leq d(a, x) + d(x, x') + d(x', b) \leq R + d(x, x') + R',$$

so $A \subset E_r(B)$ and $B \subset E_r(A)$ where $r = R + d(x, x') + R'$.

To get used to the notation, verify the following:

$$d_{\mathrm{H}}(A, B) < r \Longrightarrow \text{ for any } a \in A \text{ there is some } b \in B \text{ such that } d(a, b) < r \tag{2.1}$$

and

$$r < s \Longrightarrow \overline{E_r(A)} \subset E_s(A). \tag{2.2}$$

Theorem 2.3. *The function d_{H} is a metric on $\mathrm{H}(X)$.*

Proof. Suppose $d_{\mathrm{H}}(A, B) = 0$. This means for any $r > 0$, $A \subset E_r(B)$ and $B \subset E_r(A)$. We will show $A \subset B$ and $B \subset A$, so $A = B$.

Fix $a \in A$. For any $n \geq 1$, $A \subset E_{1/n}(B)$, so $d(a, b_n) \leq 1/n$ for some $b_n \in B$. Then as $n \to \infty$, $b_n \to a$ in X. Since B is *closed*, this implies $a \in B$. As a was arbitrary in A, we get $A \subset B$. The argument that $B \subset A$ is similar.

That d_{H} is symmetric is clear from its definition.

It remains to show the triangle inequality for d_{H}. Pick nonempty closed and bounded subsets A, B, C in X. We want to show

$$d_{\mathrm{H}}(A, B) \leq d_{\mathrm{H}}(A, C) + d_{\mathrm{H}}(C, B).$$

Let $r = d_{\mathrm{H}}(A, C)$ and $s = d_{\mathrm{H}}(C, B)$. We will show for all $\varepsilon > 0$ that

$$d_{\mathrm{H}}(A, B) \leq r + s + \varepsilon, \tag{2.3}$$

from which the triangle inequality $d_{\mathrm{H}}(A, B) \leq r + s$ follows.

To prove (2.3), pick $a \in A$. Since $d_{\mathrm{H}}(A, C) < r + \varepsilon/2$, by (2.1) there is some $c \in C$ such that $d(a, c) < r + \varepsilon/2$. Since $d_{\mathrm{H}}(C, B) < s + \varepsilon/2$, again by (2.1) there is some $b \in B$ such that $d(c, b) < s + \varepsilon/2$. Therefore

$$d(a, b) \leq d(a, c) + d(c, b) < r + s + \varepsilon,$$

so each point of A has distance less than $r+s+\varepsilon$ from some point of B. This shows $A \subset E_{r+s+\varepsilon}(B)$. Similar reasoning shows $B \subset E_{s+r+\varepsilon}(A) = E_{r+s+\varepsilon}(A)$, so (2.3) holds. $\square$

Remark 2.4. The function d_{H} makes sense on bounded subsets of X, whether or not they are closed. But if we allow bounded subsets which are not closed then we don't get a metric: any bounded subset A and its closure $\overline{A}$ satisfy $d_{\mathrm{H}}(A, \overline{A}) = 0$.

Definition 2.5. The function d_{H} is called the *Hausdorff metric* on $\mathrm{H}(X)$.

Example 2.6. If $A = \{a_0\}$ is a one-element subset of X then for any $B \in \mathrm{H}(X)$

$$\begin{aligned} d_{\mathrm{H}}(\{a_0\}, B) &= \inf\{r \geq 0 : \{a_0\} \subset E_r(B) \text{ and } B \subset E_r(\{a_0\})\} \\ &= \inf\{r \geq 0 : a_0 \in E_r(B) \text{ and } B \subset \overline{B}_r(a_0)\} \\ &= \inf\{r \geq 0 : B \subset \overline{B}_r(a_0)\} \\ &= \inf\{r \geq 0 : d(a_0, b) \leq r \text{ for all } b \in B\} \\ &= \sup_{b \in B} d(a_0, b), \end{aligned}$$

In particular, if $B = \{b_0\}$ is also a one-element set then $d_{\mathrm{H}}(\{a_0\}, \{b_0\}) = d(a_0, b_0)$.

Remark 2.7. In metric spaces, there is a useful notion of distance between a point x and a subset S, defined by

$$\operatorname{dist}(x, S) = \inf_{y \in S} d(x, y).$$

The concept makes sense for all (nonempty) subsets S, bounded or not. For instance, in $\mathbf{R}^n$ this distance between a point and a hyperplane is the length of the line segment obtained by dropping a perpendicular from the point to the hyperplane. This has a vivid geometric meaning in Euclidean space. However, this is *not* a special case of the Hausdorff distance: usually when S is closed and bounded, $\operatorname{dist}(x, S) \neq d_{\mathrm{H}}(\{x\}, S)$. Indeed, from Example 2.6 we have $d_{\mathrm{H}}(\{x\}, S) = \sup_{y \in S} d(x, y)$ rather than $\inf_{y \in S} d(x, y)$. Nevertheless, there is a connection between this notion of distance between a point and a subset and the Hausdorff metric, as we will see at the end of this section (Lemma 2.27).

Since one-point sets in X are closed and bounded, we can embed X into $\mathrm{H}(X)$ by associating to x the set $\{x\} \in \mathrm{H}(X)$. How does X look as a subset of $\mathrm{H}(X)$?

Theorem 2.8. *The function $X \to \mathrm{H}(X)$ given by $x \mapsto \{x\}$ is an isometry and the image is closed.*

Proof. The calculation in Example 2.6, with A and B both one-element sets, shows the map is an isometry. It is clearly one-to-one (in fact, any isometry is one-to-one).

Now we show the image is closed. Let $\{x_n\}$ be a sequence in X and suppose in $\mathrm{H}(X)$ that $\{x_n\} \to L$. We want to show L has only one element. (As a member of $\mathrm{H}(X)$, we at least know L is nonempty.) Suppose L has two elements, y and y'. Pick $\varepsilon < d(y, y')/2$. Since $d_{\mathrm{H}}(\{x_n\}, L) \to 0$, there is some n such that $L \subset E_\varepsilon(\{x_n\}) = \overline{B}_\varepsilon(x_n)$, so

$$d(y, y') \leq d(y, x_n) + d(x_n, y') \leq 2\varepsilon < d(y, y'),$$

a contradiction. $\square$

Theorem 2.9. *The metric space (X, d) is complete if and only if $(\mathrm{H}(X), d_{\mathrm{H}})$ is complete.*

Proof. By Theorem 2.8, X is isometric to a closed subset of $\mathrm{H}(X)$, so if the latter is complete then so is the former.

Now we will prove that if X is complete then so is $\mathrm{H}(X)$. Our argument is taken from [1, pp. 35–37], with some modifications. Let $\{A_n\}$ be a Cauchy sequence in $\mathrm{H}(X)$. Set A to be the limits of sequences $\{a_{n_1}, a_{n_2}, a_{n_3}, \dots\}$ with $n_1 < n_2 < n_3 < \cdots$ and $a_{n_i} \in A_{n_i}$ for all i:

$$A = \{a \in X : a = \lim_{i\to\infty} a_{n_i}, a_{n_i} \in A_{n_i}, n_1 < n_2 < n_3 < \cdots\}.$$

We will show in succession that

(1) $A \neq \emptyset$,
(2) A is a closed subset of X,
(3) A is a bounded subset of X (so $A \in \mathrm{H}(X)$),
(4) $d_{\mathrm{H}}(A_n, A) \to 0$ as $n \to \infty$.

1) Since $\{A_n\}$ is Cauchy in $\mathrm{H}(X)$, for every $i \geq 1$ there is an integer $n_i \geq 1$ such that $d_{\mathrm{H}}(A_m, A_n) \leq 1/2^i$ for $m, n \geq n_i$. Without loss of generality $n_1 < n_2 < n_3 < \cdots$. Now pick $a_{n_i} \in A_{n_i}$ as follows. Choose $a_{n_1} \in A_{n_1}$ arbitrarily. Once we have a_{n_i} for some i, then from $d_{\mathrm{H}}(A_{n_i}, A_{n_{i+1}}) < 1/2^i$ there is an $a_{n_{i+1}} \in A_{n_{i+1}}$ such that $d(a_{n_i}, a_{n_{i+1}}) < 1/2^i$ by (2.1). Therefore $\{a_{n_i}\}$ is a sequence getting consecutively close at a geometric rate, so it is a Cauchy sequence in X. It has a limit in X, which is an element of A, so $A \neq \emptyset$.

2) Let $\{x_n\} \subset A$ with $x_n \to x \in X$. We want to show $x \in A$. For each $i \geq 1$, pick n_i so that $d(x_{n_i}, x) < 1/i$. From the definition of A, infinitely many of the A_n's contain an element with distance less than $1/i$ from x_{n_i}. Therefore we can choose $m_1 < m_2 < m_3 < \cdots$ such that $a_{m_i} \in A_{m_i}$ and

$$d(a_{m_i}, x_{n_i}) < \frac{1}{i}.$$

Then $d(a_{m_i,n_i}, x) < 2/i$ for all i, so $x \in A$.

3) There is an $N \geq 1$ such that $d_{\mathrm{H}}(A_m, A_n) < 1$ for all $m, n \geq N$. Then $d_{\mathrm{H}}(A_m, A_N) < 1$ for $m \geq N$, so

$$A_m \subset E_1(A_N) = \bigcup_{a \in A_N} \overline{B}_1(a).$$

Since A_N is bounded, $A_N \subset \overline{B}_r(x)$ for some $x \in X$ and $r \geq 0$. Then for $m \geq N$,

$$A_m \subset \bigcup_{a \in A_N} \overline{B}_1(a) \subset \overline{B}_{r+1}(x).$$

The right side is independent of m. Taking limits of sequences $\{a_{m_i}\}$ with $a_{m_i} \in A_{m_i}$, we see that $A \subset \overline{B}_{r+1}(x)$.

4) Pick $\varepsilon > 0$. There is an $N_\varepsilon \geq 1$ such that $d_{\mathrm{H}}(A_m, A_n) < \varepsilon/2$ for all m and $n \geq N_\varepsilon$. Therefore

$$A_m \subset E_{\varepsilon/2}(A_n). \tag{2.4}$$

We will show for $n \geq N_\varepsilon$ that $A \subset E_\varepsilon(A_n)$ and $A_n \subset E_\varepsilon(A)$, so $d_{\mathrm{H}}(A, A_n) \leq \varepsilon$.

To show $A \subset E_\varepsilon(A_n)$ for $n \geq N_\varepsilon$, pick $a \in A$. Then $a = \lim_{i\to\infty} a_{m_i}$ with $a_{m_i} \in A_{m_i}$ for all i. For some $m_i \geq N_\varepsilon$, $d(a, a_{m_i}) \leq \varepsilon/2$. Since $a_{m_i} \in A_{m_i}$, by (2.4) for any $n \geq N_\varepsilon$ there is a $b_n \in A_n$ such that $d(a_{m_i}, b_n) \leq \varepsilon/2$. Therefore by the triangle inequality, for $n \geq N_\varepsilon$ we have $d(a, b_n) \leq \varepsilon$, so $a \in E_\varepsilon(A_n)$. Since $a \in A$ was arbitrary, $A \subset E_\varepsilon(A_n)$ for all $n \geq N_\varepsilon$.

Next we show $A_n \subset E_\varepsilon(A)$ for $n \geq N_\varepsilon$. Pick $x \in A_n$. We want $d(x,a) \leq \varepsilon$ for some $a \in A$. Since the sequence $A_1, A_2, A_3, \dots$ is Cauchy, we can recursively pick $n_1 < n_2 < n_3 < \cdots$ such that

$$m, m' \geq n_i \Longrightarrow d_{\mathrm{H}}(A_m, A_{m'}) < \frac{\varepsilon}{2^i}.$$

Since $n \geq N_\varepsilon$, we can use $n_1 = n$. Set $x_{n_1} = x$. If we have $x_{n_i} \in A_{n_i}$ for some i then choose $x_{n_{i+1}} \in A_{n_{i+1}}$ such that $d(x_{n_i}, x_{n_{i+1}}) < \varepsilon/2^i$. Therefore the x_{n_i}'s are Cauchy, so they converge. Call the limit a, so $a \in A$. We will show $d(x,a) \leq \varepsilon$.

From the triangle inequality,

$$d(x_{n_1}, x_{n_i}) \leq d(x_{n_1}, x_{n_2}) + d(x_{n_2}, x_{n_3}) + \cdots + d(x_{n_{i-1}}, x_{n_i}) < \frac{\varepsilon}{2} + \cdots + \frac{\varepsilon}{2^{i-1}} < \varepsilon.$$

Let $i \to \infty$: $d(x_{n_1}, a) \leq \varepsilon$. Since $x_{n_1} = x$, we're done. □

The limit of a Cauchy sequence $\{A_n\}$ in $\mathrm{H}(X)$ can be described in a simpler way.

Theorem 2.10. *Let X be a metric space and let $\{A_n\}$ be a Cauchy sequence in $\mathrm{H}(X)$. If $x = \lim_{i\to\infty} a_{n_i}$ where $n_1 < n_2 < n_3 < \cdots$ and $a_{n_i} \in A_{n_i}$, then we can fill out $\{a_{n_i}\}$ to a sequence $\{a_n\}$ with $a_n \in A_n$ for all n such that $x = \lim_{n\to\infty} a_n$. In particular, when X is complete the limit of the A_n's in $\mathrm{H}(X)$ is*

$$\{a \in X : a = \lim_{n\to\infty} a_n, a_n \in A_n \text{ for all } n\}.$$

Proof. If $n < n_1$, pick $a_n \in A_n$ arbitrarily. If $n_i < n < n_{i+1}$ for some i we want to pick $a_n \in A_n$ to be close to a_{n_i}. Using (2.1), we can pick $a_n \in A_n$ such that $d(a_n, a_{n_i}) < d_{\mathrm{H}}(A_n, A_{n_i}) + 1/n$. This constructs a sequence $\{a_n\}$ with $a_n \in A_n$ for all n.

To prove $d(a_n, x) \to 0$ as $n \to \infty$, pick $\varepsilon > 0$. Since $a_{n_i} \to x$, there is i_0 such that for $i \geq i_0$, $d(a_{n_i}, x) \leq \varepsilon$. Since the A_n's are Cauchy in $\mathrm{H}(X)$, there is an $N \geq 1$ such that $d_{\mathrm{H}}(A_m, A_n) \leq \varepsilon$ for $m, n \geq N$. Increase N more if necessary so we also have $1/N \leq \varepsilon$. Let $M = \max(n_{i_0}, N)$. We will show $d(a_n, x) \leq \varepsilon$ whenever $n \geq M$. For such n, there is an i such that $n_i \leq n < n_{i+1}$. Then $i \geq i_0$ since $M \geq n_{i_0}$. If $n = n_i$ then we have $d(a_n, x) \leq \varepsilon$ from the choice of i_0. If $n_i < n < n_{i+1}$, then

$$\begin{aligned} d(a_n, x) &\leq d(a_n, a_{n_i}) + d(a_{n_i}, x) \\ &< d_{\mathrm{H}}(A_n, A_{n_i}) + \frac{1}{n} + \varepsilon \quad \text{by the choice of } a_n \\ &\leq \varepsilon + \frac{1}{N} + \varepsilon \quad \text{since } n, n_i \geq N \\ &\leq 3\varepsilon. \end{aligned}$$

That proves $a_n \to x$.

The description of the the limit of the A_n's, when X is complete, as limits of sequences in X with one term from each A_n, follows from Theorem 2.9 and the filling out procedure we just described. □

Remark 2.11. In Theorem 2.9 we could have described the limit of the A_n's in the less complicated way from Theorem 2.10. However, it was actually sequences $\{a_{n_i}\}$ which showed up in the proof of Theorem 2.9, rather than sequences $\{a_n\}$ with $a_n \in A_n$ for all n, so we would have had to prove Theorem 2.10 before Theorem 2.9 and then invoke it 3 times (do you see where?) in the proof of Theorem 2.9. That would make the proof of Theorem 2.9 seem more complicated than it really is. It seems simpler to use the less tidy description of $\lim A_n$ first and then come back later to fix it up.

The set $\mathrm{H}(X)$ is closed under finite unions. The next theorem estimates the effect of unions on the Hausdorff distance.

Theorem 2.12. *For subsets $A_1, \dots, A_m, B_1, \dots, B_m$ of $\mathrm{H}(X)$,*

$$d_{\mathrm{H}}(A_1 \cup \dots \cup A_m, B_1 \cup \dots \cup B_m) \le \max_{1 \le i \le m} d_{\mathrm{H}}(A_i, B_i).$$

Proof. We will check the case $m = 2$. The rest is a simple induction.

Pick $A, B, C, D \in \mathrm{H}(X)$. We want to show $d_{\mathrm{H}}(A \cup B, C \cup D) \le \max(d_{\mathrm{H}}(A, C), d_{\mathrm{H}}(B, D))$.

Let $r = d_{\mathrm{H}}(A, C)$ and $s = d_{\mathrm{H}}(B, D)$. Without loss of generality, $r \le s$. So we want to show $d_{\mathrm{H}}(A \cup B, C \cup D) \le s$. We will show for any $\varepsilon > 0$ that $d_{\mathrm{H}}(A \cup B, C \cup D) \le s + \varepsilon$.

Pick $x \in A$. Since $d_{\mathrm{H}}(A, C) < r + \varepsilon$, there is a $y \in C$ such that $d(x, y) < r + \varepsilon \le s + \varepsilon$. Therefore $A \subset E_{s+\varepsilon}(C) \subset E_{s+\varepsilon}(C \cup D)$. Similarly, $B \subset E_{s+\varepsilon}(C \cup D)$, so $A \cup B \subset E_{s+\varepsilon}(C \cup D)$. The inclusion $C \cup D \subset E_{s+\varepsilon}(A \cup B)$ is proved in a similar way. Therefore $d_{\mathrm{H}}(A \cup B, C \cup D) \le s + \varepsilon$. □

If $f\colon X \to X$ is a contraction with constant c then for any closed ball $\overline{B}_r(x)$ in X we have $f(\overline{B}_r(x)) \subset \overline{B}_{cr}(f(x))$, so f sends bounded sets to bounded sets. The closure of a bounded set is bounded, so we get a function $\mathrm{H}(X) \mapsto \mathrm{H}(X)$ by $A \mapsto \overline{f(A)}$.

Theorem 2.13. *Let $f\colon X \to X$ be a contraction with constant c. Then the induced map on $\mathrm{H}(X)$ given by $A \mapsto \overline{f(A)}$ is a contraction with the same constant: $d_{\mathrm{H}}(\overline{f(A)}, \overline{f(B)}) \le c d_{\mathrm{H}}(A, B)$ for all $A, B \in \mathrm{H}(X)$.*

Proof. If $c = 0$ then f is constant with image $\{x_0\}$, say, so $\overline{f(A)} = \{x_0\}$ for all $A \in \mathrm{H}(X)$.

Now take $c > 0$. Set $r = d_{\mathrm{H}}(A, B)$. To show $d_{\mathrm{H}}(\overline{f(A)}, \overline{f(B)}) \le cr$, we will show

$$d_{\mathrm{H}}(\overline{f(A)}, \overline{f(B)}) \le c(r + \varepsilon)$$

for all $\varepsilon > 0$. This will follow from showing

$$\overline{f(A)} \subset E_{c(r+\varepsilon)}(\overline{f(B)}), \quad \overline{f(B)} \subset E_{c(r+\varepsilon)}(\overline{f(A)}).$$

Pick $a \in A$. Since $d_{\mathrm{H}}(A, B) < r + \varepsilon/2$, by (2.1) there is a $b \in B$ such that $d(a, b) < r + \varepsilon/2$, so $d(f(a), f(b)) \le c d(a, b) < c(r + \varepsilon/2)$. Thus $f(a) \in E_{c(r+\varepsilon/2)}(f(B)) \subset E_{c(r+\varepsilon/2)}(\overline{f(B)})$. Since a was arbitrary in A,

$$f(A) \subset E_{c(r+\varepsilon/2)}(\overline{f(B)}).$$

Taking the closure of both sides,

$$\overline{f(A)} \subset \overline{E_{c(r+\varepsilon/2)}(\overline{f(B)})}.$$

This inclusion implies, by (2.2),

$$\overline{f(A)} \subset E_{c(r+\varepsilon)}(\overline{f(B)}).$$

Similar reasoning gives

$$\overline{f(B)} \subset E_{c(r+\varepsilon)}(\overline{f(A)}).$$

□

Theorem 2.14 (Hutchinson). *Let $f_1, \dots, f_m\colon X \to X$ be contractions. Set $F\colon \mathrm{H}(X) \to \mathrm{H}(X)$ by*

$$F(A) = \overline{f_1(A)} \cup \dots \cup \overline{f_m(A)}.$$

If f_i has contraction constant c_i, F is a contraction with constant $\max(c_1, \dots, c_m)$.

Proof. Using Theorems 2.12 and 2.13,

$$d_{\mathrm{H}}(F(A), F(B)) = d_{\mathrm{H}}\left(\bigcup_{i=1}^{m} \overline{f_i(A)}, \bigcup_{i=1}^{m} \overline{f_i(B)}\right) \leq \max_{1\leq i\leq m} d_{\mathrm{H}}(\overline{f_i(A)}, \overline{f_i(B)}) \leq \max_{1\leq i\leq m} c_i d_{\mathrm{H}}(A, B).$$

□

Corollary 2.15. *If X is a complete metric space and $f_1, \ldots, f_m$ are contractions on X, there is a unique nonempty closed and bounded subset $A \subset X$ such that*

$$\overline{f_1(A)} \cup \cdots \cup \overline{f_m(A)} = A.$$

Proof. Since X is complete, $\mathrm{H}(X)$ is complete. The function F as in Theorem 2.14 is a contraction on $\mathrm{H}(X)$, so by the contraction mapping theorem on $\mathrm{H}(X)$ there is a unique $A \in \mathrm{H}(X)$ such that $F(A) = A$. □

When $m = 1$, Corollary 2.15 is dull: letting a be the fixed point of f_1, we have $F(\{a\}) = \overline{\{f_1(a)\}} = \overline{\{a\}} = \{a\}$, so $\{a\}$ is the fixed point of F in $\mathrm{H}(X)$. Taking $m > 1$, things get much more interesting. With $X = \mathbf{R}^n$, we will see that some fixed points in this setting are fractals. Since $\mathrm{H}(\mathbf{R}^n)$ is the set of nonempty compact subsets of $\mathbf{R}^n$, we do not need to take closures in the definition of F because each $f_i(A)$ is already compact (and thus closed):

$$F(A) = \overline{f_1(A)} \cup \cdots \cup \overline{f_m(A)} = f_1(A) \cup \cdots \cup f_m(A).$$

We may write $F = f_1 \cup \cdots \cup f_m$ to describe this function. Its iterates are computed by composing the f_i's in all possible ways and then taking the union, *e.g.*, if $F = f_1 \cup f_2$ then $F^2 = f_1^2 \cup (f_1 \circ f_2) \cup (f_2 \circ f_1) \cup f_2^2$.

Example 2.16. Take $X = \mathbf{R}$. Let $f_0, f_1 \colon \mathbf{R} \to \mathbf{R}$ by

$$f_0(x) = \frac{1}{3}x, \quad f_1(x) = \frac{1}{3}(x-1) + 1 = \frac{1}{3}x + \frac{2}{3}.$$

Both f_0 and f_1 are contractions on $\mathbf{R}$, so by Corollary 2.15 there is a unique nonempty compact subset $A \subset \mathbf{R}$ such that $A = f_0(A) \cup f_1(A)$. This is a self-similarity property: A is a union of two copies of itself, each at one-third the original size. This is one of the features of the Cantor set, so A is the Cantor set because there is only one compact set fixed by $f_0 \cup f_1$.

The Cantor set was first defined by Cantor (1884) using the recipe of removing successive middle-thirds starting with $[0, 1]$, but the viewpoint of Corollary 2.15 shows that the Cantor set can be obtained from any compact set $C \subset \mathbf{R}$: the sequence $C, F(C), F(F(C)), \ldots$, where $F(C) = f_0(C) \cup f_1(C)$, converges in $\mathrm{H}(\mathbf{R})$ to the Cantor set. Taking $C = [0, 1]$, the iterates $F^n(C)$ are precisely the result of removing successive middle-thirds. But the convergence to the Cantor set using F-iterates can be achieved starting with any (nonempty) compact subset $C \subset \mathbf{R}$, even a single point!

Example 2.17. Use $X = \mathbf{R}$ as before but now take $f_0(x) = x/2$ and $f_1(x) = (x-1)/2+1 = (x+1)/2$. Check that $A = [0, 1]$ satisfies $A = f_0(A) \cup f_1(A)$, so $[0, 1]$ is the unique compact fixed point (and is not really a fractal). There are non-compact fixed points of $f_0 \cup f_1$ in $\mathrm{H}(\mathbf{R})$, such as $A = \mathbf{R}$ and $A = (0, 1)$. Only the compact fixed point in $\mathrm{H}(\mathbf{R})$ is unique.

Example 2.18. Take $X = \mathbf{R}^2$. For any point p in the plane, let $f_p \colon \mathbf{R}^2 \to \mathbf{R}^2$ be the function which sends a point q to the point halfway closer to p along the segment connecting p and q. (In particular, $f_p(p) = p$.) Easily f_p is a contraction in the usual metric on $\mathbf{R}^2$, with contraction constant $1/2$.

Pick 3 different points in the plane, say u, v, and w. Then the three functions f_u, f_v, and f_w are contractions on $\mathbf{R}^2$ with contraction constant $1/2$, so $F = f_u \cup f_v \cup f_w$ is a contraction mapping on $\mathrm{H}(\mathbf{R}^2)$ with $d_{\mathrm{H}}(F(A), F(A')) \leq (1/2)d_{\mathrm{H}}(A, A')$ for all A and A' in $\mathrm{H}(\mathbf{R}^2)$. There is a unique nonempty compact subset $A \subset \mathbf{R}^2$ satisfying

$$A = f_u(A) \cup f_v(A) \cup f_w(A).$$

The Sierpinski triangle with vertices u, v, and w satisfies this condition, so A is that Sierpinski triangle.

Example 2.19. Viewing $\mathbf{R}^2$ as the complex plane, let $f_1, f_2, f_3, f_4 \colon \mathbf{C} \to \mathbf{C}$ by

$$f_1(z) = \frac{1}{3}z, \quad f_2(z) = \frac{1}{3}e^{i\pi/6}z + \frac{1}{3}, \quad f_3(z) = \frac{1}{3}e^{-i\pi/6}z + \left(\frac{1}{3}e^{i\pi/6} + \frac{1}{3}\right), \quad f_4(z) = \frac{1}{3}z + \frac{2}{3}.$$

The (nonempty) compact $A \subset \mathbf{C}$ satisfying

$$A = f_1(A) \cup f_2(A) \cup f_3(A) \cup f_4(A)$$

is the Koch curve fractal with endpoints 0 and 1.

A finite collection of contraction mappings on a complete metric space X is called an *iterated function system*. We have seen that it leads to a single contraction mapping on $\mathrm{H}(X)$, whose fixed point in $\mathrm{H}(X)$ includes some classical fractals when X is Euclidean space. For an application of iterated function systems to the construction of a continuous nowhere differentiable function, see [11]. Applications of iterated function systems to computer graphics are discussed in [1] and [9].

The fixed point of an iterated function system on $\mathbf{R}^n$ is compact since all members of $\mathrm{H}(\mathbf{R}^n)$ are compact. For a general complete metric space X, there can be closed and bounded subsets which are not compact. But it is *still* true that the fixed point of an iterated function system on X is compact. To prove this, we introduce

$$\mathrm{K}(X) = \text{nonempty compact subsets of } X.$$

Since compact subsets of X are closed and bounded, $\mathrm{K}(X) \subset \mathrm{H}(X)$.

Theorem 2.20. *If X is complete then $\mathrm{K}(X)$ is a closed subset of $\mathrm{H}(X)$. If X is incomplete then $\mathrm{K}(X)$ is not closed in $\mathrm{H}(X)$.*

Proof. First we suppose X is complete. Let $\{A_n\}$ be a sequence in $\mathrm{K}(X)$ which converges to A in $\mathrm{H}(X)$. We want to show A is a compact subset of X. We will use a compactness criterion for general metric spaces: a subset of a metric space is compact if and only if the subset is complete and totally bounded. (Totally bounded means for any $\varepsilon > 0$ that the covering by open ε-balls has a finite subcovering.)

Since A is a closed subset of X and X is complete, A is complete. It remains to show A is a totally bounded subset of X. Pick any $\varepsilon > 0$. If A is *not* covered by finitely many open ε-balls then there is a sequence $\{x_n\}$ in A such that $d(x_m, x_n) \geq \varepsilon$ for all $m \neq n$. Since $A_n \to A$ in $\mathrm{H}(X)$, for some $N \geq 1$ we have $A \subset E_{\varepsilon/3}(A_N)$. Therefore, for any n, there is $y_n \in A_N$ such that $d(x_n, y_n) \leq \varepsilon/3$. We now have a sequence $\{y_n\}$ in A_N. Since A_N is compact, the y_n's have a convergent subsequence. So for *some* pair of (large) distinct integers m and n, $d(y_m, y_n) < \varepsilon/3$. Then by the triangle inequality,

$$d(x_m, x_n) \leq d(x_m, y_m) + d(y_m, y_n) + d(y_n, x_n) < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon,$$

a contradiction.

Now suppose X is incomplete: there is a Cauchy sequence $\{x_1, x_2, \dots\}$ in X without a limit in X. We will use this non-convergent Cauchy sequence to write down a sequence in $\mathrm{K}(X)$ which converges to a set in $\mathrm{H}(X) - \mathrm{K}(X)$. Let $A_n = \{x_1, \dots, x_n\}$ and $A = \{x_1, x_2, \dots\}$. Each A_n is finite, so also compact: $A_n \in \mathrm{K}(X)$ for all n. We will show $A \in \mathrm{H}(X)$, $A \notin \mathrm{K}(X)$, and $A_n \to A$ in $\mathrm{H}(X)$.

Since A has no limit points in X (if a subsequence of a Cauchy sequence converges then the Cauchy sequence itself converges), A is a closed subset of X. Since Cauchy sequences are bounded, A is a bounded set. Therefore $A \in \mathrm{H}(X)$. The set A is not compact because none of its subsequences converge in X, so $A \notin \mathrm{K}(X)$. To show $d_{\mathrm{H}}(A_n, A) \to 0$ as $n \to \infty$, it suffices to show for any $\varepsilon > 0$ that $A \subset E_\varepsilon(A_n)$ for $n \gg 0$; the inclusion $A_n \subset E_\varepsilon(A)$ will be automatic because $A_n \subset A$.

Pick $\varepsilon > 0$. There is some $N \geq 1$ such that $d(x_m, x_n) \leq \varepsilon$ for all $m, n \geq N$. Therefore $x_m \in E_\varepsilon(A_n)$ for all $m, n \geq N$. Since $x_1, \dots, x_{N-1} \in A_n$ when $n \geq N$ we get $A \subset E_\varepsilon(A_n)$ for $n \geq N$. □

Here is the analogue for $\mathrm{K}(X)$ of Theorem 2.9.

Corollary 2.21. *The metric space X is complete if and only if $\mathrm{K}(X)$ is complete in the Hausdorff metric.*

Proof. If X is complete then so is $\mathrm{H}(X)$. A closed subset of a complete space is complete, so $\mathrm{K}(X)$ is complete in the Hausdorff metric by Theorem 2.20.

Conversely, suppose $\mathrm{K}(X)$ is complete. The isometric embedding $X \to \mathrm{H}(X)$ which sends each x to $\{x\}$ has image in $\mathrm{K}(X)$, and the proof of Theorem 2.8 carries over to show the image is a closed subset of $\mathrm{K}(X)$ (not just closed in $\mathrm{H}(X)$), so X is complete. □

Lemma 2.22. *Let $f\colon X \to X$ be a contraction on a complete metric space and $Y \subset X$ be a closed subset such that $f(Y) \subset Y$. Then the unique fixed point of f is in Y.*

Proof. Since Y is a closed subset of a complete metric space, it is complete. Then we can apply the contraction mapping theorem to $f\colon Y \to Y$, so f has a fixed point in Y. Since f has only one fixed point in X, it must lie in Y. □

Theorem 2.23. *Let X be complete and $\{f_1, \dots, f_m\}$ be an iterated function system on X with associated contraction F on $\mathrm{H}(X)$. The unique fixed point of F in $\mathrm{H}(X)$ is compact.*

Proof. For $A \in \mathrm{K}(X)$, each $f_i(A)$ is compact, so $F(A)$ is a union of finitely many compact sets and therefore is compact. Hence $F\colon \mathrm{K}(X) \to \mathrm{K}(X)$, so the unique fixed point of F in $\mathrm{H}(X)$ must lie in $\mathrm{K}(X)$ by Lemma 2.22. □

Theorem 2.23 explains why some treatments of iterated function systems, such as [1], focus on $\mathrm{K}(X)$ rather than $\mathrm{H}(X)$ right from the start.

Theorem 2.24. *The metric space X is compact if and only if $\mathrm{H}(X)$ is compact.*

Proof. Since X is isometric to a closed subset of $\mathrm{H}(X)$, if $\mathrm{H}(X)$ is compact then so is X.

Conversely, suppose X is compact. To show $\mathrm{H}(X)$ is compact we will show it is complete and totally bounded. From completeness of X we have completeness of $\mathrm{H}(X)$. To show $\mathrm{H}(X)$ is totally bounded, pick $\varepsilon > 0$. Then X is covered by finitely many open $\varepsilon/2$-balls, say

$$X = B_{\varepsilon/2}(x_1) \cup B_{\varepsilon/2}(x_2) \cup \cdots \cup B_{\varepsilon/2}(x_n).$$

We will show $\mathrm{H}(X)$ has a finite covering by open ε-balls in the Hausdorff metric, namely the open ε-balls in $\mathrm{H}(X)$ around the finitely many nonempty subsets of $\{x_1, \ldots, x_n\}$. (Note finite sets are compact.)

Pick any $A \in \mathrm{H}(X)$. It has to meet some of the $B_{\varepsilon/2}(x_i)$'s, since those balls cover X. Let S be the set of x_i's with that property:

$$S = \{x_i : A \cap B_{\varepsilon/2}(x_i) \neq \emptyset\}.$$

Every element of A is within distance $\varepsilon/2$ of some x_i in S, so $A \subset E_{\varepsilon/2}(S)$. Each $x_i \in S$ is within $\varepsilon/2$ of some element of A, so $S \subset E_{\varepsilon/2}(A)$. Thus $d_{\mathrm{H}}(S, A) \leq \varepsilon/2 < \varepsilon$. Since A was arbitrary, the open ε-balls in $\mathrm{H}(X)$ around the nonempty subsets S of $\{x_1, \ldots, x_n\}$ cover $\mathrm{H}(X)$. □

Lemma 2.25. *Let X be a compact metric space. If $f \colon X \to X$ satisfies $d(f(x), f(x')) < d(x, x')$ when $x \neq x'$ in X, then f has a unique fixed point in X and the fixed point can be found as the limit of $f^n(x_0)$ as $n \to \infty$ for any $x_0 \in X$.*

This is due to Edelstein [6].

Proof. To show f has at most one fixed point in X, suppose f has two fixed points $a \neq a'$. Then $d(a, a') = d(f(a), f(a')) < d(a, a')$. This is impossible, so $a = a'$.

To prove f actually has a fixed point, we will look at the function $X \to [0, \infty)$ given by $x \mapsto d(x, f(x))$. This measures the distance between each point and its f-value. A fixed point of f is where this function takes the value 0.

Since X is compact, the function $d(x, f(x))$ takes on its minimum value: there is an $a \in X$ such that $d(a, f(a)) \leq d(x, f(x))$ for all $x \in X$. We'll show by contradiction that a is a fixed point for f. If $f(a) \neq a$ then the hypothesis about f in the theorem (taking $x = a$ and $x' = f(a)$) says

$$d(f(a), f(f(a))) < d(a, f(a)),$$

which contradicts the minimality of $d(a, f(a))$ among all numbers $d(x, f(x))$. So $f(a) = a$.

Finally, we show for any $x_0 \in X$ that the sequence $x_n = f^n(x_0)$ converges to a as $n \to \infty$. This can't be done as in the proof of the contraction mapping theorem since we don't have the contraction constant to help us out. Instead we will exploit compactness.

If for some $k \geq 0$ we have $x_k = a$ then $x_{k+1} = f(x_k) = f(a) = a$, and more generally $x_n = a$ for all $n \geq k$, so $x_n \to a$ since the terms of the sequence equal a for all large n. Now we may assume instead that $x_n \neq a$ for all n. Then

$$0 < d(x_{n+1}, a) = d(f(x_n), f(a)) < d(x_n, a),$$

so the sequence of numbers $d(x_n, a)$ is decreasing and positive. Thus it has a limit $\ell = \lim_{n\to\infty} d(x_n, a) \geq 0$. We will show $\ell = 0$ (so $d(x_n, a) \to 0$, which means $x_n \to a$ in X). By compactness of X, the sequence $\{x_n\}$ has a convergent subsequence x_{n_i}, say $x_{n_i} \to y \in X$. Then, by continuity of f, $f(x_{n_i}) \to f(y)$, which says $x_{n_i+1} \to f(y)$ as $i \to \infty$. Since $d(x_n, a) \to \ell$ as $n \to \infty$, $d(x_{n_i}, a) \to \ell$ and $d(x_{n_i+1}, a) \to \ell$ as $i \to \infty$. By continuity of the metric, $d(x_{n_i}, a) \to d(y, a)$ and $d(x_{n_i+1}, a) = d(f(x_{n_i}), a) \to d(f(y), a)$. Having already shown these limits are ℓ,

$$d(y, a) = \ell = d(f(y), a) = d(f(y), f(a)). \tag{2.5}$$

If $y \neq a$ then $d(f(y), f(a)) < d(y, a)$, but this contradicts (2.5). So $y = a$, which means $\ell = d(y, a) = 0$. That shows $d(x_n, a) \to 0$ as $n \to \infty$. □

Lemma 2.25 is a fixed-point theorem for a function f on a compact metric space satisfying $d(f(x), f(y)) < d(x, y)$ when $x \neq y$. For such a function f there is an analogue of Theorem 2.13, as follows.

Theorem 2.26. *Let $f\colon X \to X$ satisfy $d(f(x), f(y)) < d(x, y)$ for all distinct x and y in X. Then for all distinct A and B in $\mathrm{K}(X)$, $d_{\mathrm{H}}(f(A), f(B)) < d_{\mathrm{H}}(A, B)$.*

To prove this, we use a technical property of the Hausdorff distance between compact subsets, codified in the following lemma.

Lemma 2.27. *For $x \in X$ and $S \subset X$, set the distance from x to S to be $\operatorname{dist}(x, S) = \inf_{y\in S} d(x, y)$. For compact subsets A and B in X, either $d_{\mathrm{H}}(A, B) = \operatorname{dist}(x, B)$ for some $x \in A$ or $d_{\mathrm{H}}(A, B) = \operatorname{dist}(x, A)$ for some $x \in B$.*

Proof. First we show for $r > 0$ that

$$A \subset E_r(B) \Longleftrightarrow \text{for all } a \in A, \operatorname{dist}(a, B) \leq r. \tag{2.6}$$

Compactness will be essential.

If $A \subset E_r(B)$ then for all $a \in A$ there is a $b \in B$ such that $d(a, b) \leq r$, so $\operatorname{dist}(a, B) \leq d(a, b) \leq r$. Conversely, suppose $\operatorname{dist}(a, B) \leq r$ for all $a \in A$. Since B is compact and $d(a, x)$ is continuous in x, its infimum over B is a value on B:

$$\operatorname{dist}(a, B) = \inf_{b\in B} d(a, b) = d(a, b')$$

for some $b' \in B$. Then $d(a, b') \leq r$, so $a \in E_r(B)$. This holds for all a in A, so $A \subset E_r(B)$. This concludes the proof of (2.6).

For any (nonempty) subset $S \subset X$, let $\delta_S\colon X \to \mathbf{R}$ by

$$\delta_S(x) = \operatorname{dist}(x, S) = \inf_{y\in S} d(x, y). \tag{2.7}$$

Then (2.6) says $A \subset E_r(B)$ if and only if $\delta_B(a) \leq r$ for all $a \in A$, which is equivalent to $\sup_{a\in A} \delta_B(a) \leq r$. Returning to (2.6), we have

$$A \subset E_r(B) \text{ and } B \subset E_r(A) \Longleftrightarrow \sup_{a\in A} \delta_B(a) \leq r \text{ and } \sup_{b\in B} \delta_A(b) \leq r,$$

so

$$\begin{aligned} d_{\mathrm{H}}(A, B) &= \inf\{r \geq 0 : \sup_{a\in A} \delta_B(a) \leq r \text{ and } \sup_{b\in B} \delta_A(b) \leq r\} \\ &= \inf\{r \geq 0 : \max(\sup_{a\in A} \delta_B(a), \sup_{b\in B} \delta_A(b)) \leq r\} \\ &= \max(\sup_{a\in A} \delta_B(a), \sup_{b\in B} \delta_A(b)). \end{aligned}$$

With this formula for $d_{\mathrm{H}}(A, B)$, the proof of the lemma boils down to showing the supremum of δ_B over A is a value of δ_B on A and the supremum of δ_A over B is a value of δ_A on B. This will follow from the continuity of δ_A and δ_B on X and the compactness of A and B (a continuous real-valued function on a compact set assumes its supremum as a value). To show δ_A and δ_B are continuous, we will show more generally that δ_S is continuous for any subset S of X. To be precise, we will show

$$|\,\delta_S(x) - \delta_S(y)| \leq d(x, y) \tag{2.8}$$

for all x and y in X. For $\varepsilon > 0$ there is an $s \in S$ such that $d(x, s) < \delta_S(x) + \varepsilon$. Then

$$\delta_S(y) \leq d(y, s) \leq d(y, x) + d(x, s) < d(y, x) + \delta_S(x) + \varepsilon,$$

so

$$\delta_S(y) - \delta_S(x) < d(y,x) + \varepsilon.$$

Now let $\varepsilon \to 0^+$:

$$\delta_S(y) - \delta_S(x) \leq d(y,x) = d(x,y).$$

Reversing the roles of x and y, we get

$$\delta_S(x) - \delta_S(y) \leq d(x,y),$$

so (2.8) follows.

By continuity of δ_A and δ_B and compactness of A and B,

$$\sup_{a\in A} \delta_B(a) = \delta_B(a_0) \text{ and } \sup_{b\in B} \delta_A(b) = \delta_A(b_0)$$

for some $a_0 \in A$ and $b_0 \in B$. Therefore

$$d_{\mathrm{H}}(A,B) = \max(\delta_B(a_0), \delta_A(b_0)) = \max(\operatorname{dist}(a_0,B), \operatorname{dist}(b_0,A)).$$

□

Now we prove Theorem 2.26.

Proof. Since $A \neq B$, $d_{\mathrm{H}}(A,B) > 0$. We want to prove $d_{\mathrm{H}}(f(A), f(B)) < d_{\mathrm{H}}(A,B)$. Therefore we may suppose $f(A) \neq f(B)$. Since f is continuous, $f(A)$ and $f(B)$ are both compact. We apply Lemma 2.27 to $f(A)$ and $f(B)$: either $d_{\mathrm{H}}(f(A), f(B)) = \operatorname{dist}(f(x), f(B))$ for some $x \in A$ or $d_{\mathrm{H}}(f(A), f(B)) = \operatorname{dist}(f(x), f(A))$ for some $x \in B$. From the symmetry in A and B, we may suppose the first formula holds, so

$$d_{\mathrm{H}}(f(A), f(B)) = \operatorname{dist}(f(x), f(B)) \leq d(f(x), f(b))$$

for some $x \in A$ and all $b \in B$. Since B is compact, $\operatorname{dist}(x,B) = d(x,y)$ for some $y \in B$, so no element of B is closer to x than y is. Therefore $d_{\mathrm{H}}(A,B) \geq d(x,y)$. (If $d_{\mathrm{H}}(A,B) < d(x,y)$ then x is within distance less than $d(x,y)$ of some element of B, a contradiction.) Now we have

$$0 < d_{\mathrm{H}}(f(A), f(B)) = \operatorname{dist}(f(x), f(B)) \leq d(f(x), f(y)).$$

Since $d(f(x), f(y)) > 0$ we have $x \neq y$, so $d(f(x), f(y)) < d(x,y) \leq d_{\mathrm{H}}(A,B)$. Thus $d_{\mathrm{H}}(f(A), f(B)) < d_{\mathrm{H}}(A,B)$. □

Corollary 2.28. *Let $f_1, \ldots, f_m \colon X \to X$ each satisfy $d(f_i(x), f_i(y)) < d(x,y)$ whenever $x \neq y$ in X. Let $F = f_1 \cup \cdots \cup f_m \colon \mathrm{K}(X) \to \mathrm{K}(X)$. Then $d_{\mathrm{H}}(F(A), F(B)) < d_{\mathrm{H}}(A,B)$ whenever $A \neq B$ in $\mathrm{K}(X)$.*

Proof. Use Theorems 2.12 and 2.26. □

Example 2.29. Let $f_1, f_2 \colon [0,1] \to [0,1]$ by $f_1(x) = 1/(1+x)$ and $f_2(x) = 1/(2+x)$. Since $|f(x)-f(y)|/|x-y| = 1/(1+x)(1+y)$, $|f(x)-f(y)|/|x-y|$ gets arbitrarily close to 1 when x and y are sufficiently close to 0, so f_1 is not a contraction, but $|f_1(x) - f_1(y)| < |x-y|$ for $x \neq y$ in $[0,1]$. Since $|f_2'(x)| = 1/(2+x)^2$, f_2 is a contraction on $[0,1]$ with constant $1/4$. Let $F = f_1 \cup f_2$ on $\mathrm{H}([0,1])$. By Corollary 2.28, $d_{\mathrm{H}}(F(A), F(B)) < d_{\mathrm{H}}(A,B)$ for all $A \neq B$ in $\mathrm{H}([0,1]) = \mathrm{K}([0,1])$.

Since $\mathrm{H}([0,1])$ is compact, by Lemma 2.25 there is a unique fixed point of F in $\mathrm{H}([0,1])$ and the sequence of iterates $C, F(C), F^2(C), \ldots$ starting from any $C \in \mathrm{H}([0,1])$ converges to this fixed point. This seems to be a serious use of Lemma 2.25, because F truly is not a contraction. Indeed, the reader should check that for distinct a and b close to 0 in $[0,1]$,

$$d_{\mathrm{H}}(F(\{a\}), F(\{b\})) = \frac{1}{(1+a)(1+b)} d_{\mathrm{H}}(\{a\}, \{b\}),$$

which in more concrete terms says

$$d_{\mathrm{H}}(\{f_1(a), f_2(a)\}, \{f_1(b), f_2(b)\}) = \frac{|a-b|}{(1+a)(1+b)}.$$

Therefore $d_{\mathrm{H}}(F(\{a\}), F(\{b\}))/d_{\mathrm{H}}(\{a\}, \{b\})$ can be made arbitrarily close to 1 by taking a and b sufficiently close to 0. (For simplicity, even take one of a or b to be 0.)

Alas, this example can be made into an application of the usual contraction mapping theorem by looking at $F^2 = f_1^2 \cup (f_1 \circ f_2) \cup (f_2 \circ f_1) \cup f_2^2$. Explicitly compute each of the rational functions $f_1^2, f_1 \circ f_2, f_2 \circ f_1$, and f_2^2 and verify they are contractions by maximizing their derivatives on $[0,1]$ (f_1^2 has contraction constant $1/4$, f_2^2 has constant $1/16$, and the two composites $f_1 \circ f_2$ and $f_2 \circ f_1$ have constant $1/9$). Therefore we can apply the contraction mapping theorem to F^2.

What is the fixed point of F, as an explicit subset of $[0,1]$? A well-chosen initial set C will give us the answer in terms of continued fractions. Take $C = \{0\}$. The set $F^n(\{0\})$ consists of the rational numbers in $[0,1]$ with a continued fraction expansion of length n containing only 1's and 2's (ignoring the initial continued fraction entry 0 common to all continued fractions in $[0,1]$): $F(\{0\}) = \{1, 1/2\}$, $F^2(\{0\}) = \{1/2, 1/3, 2/3, 2/5\}$, and so on. The fixed point of F in $\mathrm{H}([0,1])$ is $\lim_{n\to\infty} F^n(\{0\})$, and by the proof of Theorem 2.9 this is the set of all limits of sequences $\{a_{n_i}\}$ where $n_1 < n_2 < \cdots$ and $a_{n_i} \in F^{n_i}(\{0\})$ for all i. These limits are precisely the real numbers in $[0,1]$ whose continued fraction expansion contains only 1's and 2's: any such continued fraction expansion is clearly a limit of such rational continued fractions, and any continued fraction expansion containing an entry besides 1 and 2 is not a limit of such rational continued fractions since a continued fraction that is sufficiently close to a given continued fraction must have the same initial entries.

Example 2.30. Following [2], let $f\colon [0,1] \to [0,1]$ by $f(x) = x/(1+x)$. Since $f^n(x) = x/(1+nx)$, for $x \neq y$ the ratio $|f^n(x) - f^n(y)|/|x-y| = 1/(1+nx)(1+ny)$ is arbitrarily close to 1 when x and y are sufficiently close to 0, so no iterate f^n is a contraction on any neighborhood of its fixed point 0.

Example 2.31. Let $f_1, f_2\colon [0,1] \to [0,1]$ by $f_1(x) = x/(1+x)$ and $f_2(x) = 1/(2-x)$. These functions satisfy $|f(x) - f(y)| < |x-y|$ for $x \neq y$ and none of their iterates are contractions on $[0,1]$. Let $F = f_1 \cup f_2$ on $\mathrm{H}([0,1])$, so $d_{\mathrm{H}}(F(A), F(B)) < d_{\mathrm{H}}(A,B)$ for $A \neq B$ in $\mathrm{H}([0,1])$ by Corollary 2.28. Since no iterate of f_1 or f_2 is a contraction on $[0,1]$, it is reasonable to expect no iterate of F is a contraction on $\mathrm{H}([0,1])$, which would make the existence of a (unique) fixed point of F in $\mathrm{H}([0,1])$ a real use of Lemma 2.25 and the compactness of $\mathrm{H}([0,1])$. The fixed point of F is easy to identify: $f_1([0,1]) = [0,1/2]$ and $f_2([0,1]) = [1/2,1]$, so $F([0,1]) = [0,1]$. Thus $[0,1]$ is the only fixed point of F in $\mathrm{H}([0,1])$.

How can we show no iterate of F is a contraction? The idea is to mimic the argument in Example 2.30: look at how F^n changes distances in the neighorhood of its fixed point $[0,1]$. We will show for each n that the ratio $d_{\mathrm{H}}(F^n(A), F^n([0,1]))/d_{\mathrm{H}}(A,[0,1]) = d_{\mathrm{H}}(F^n(A),[0,1])/d_{\mathrm{H}}(A,[0,1])$ can be made arbitrarily close to 1 by a suitable choice of A, so F^n is not a contraction.

Using the definition of the Hausdorff distance, $d_{\rm H}(A,[0,1]) = \inf\{r \geq 0 : [0,1] \subset E_r(A)\}$ (since $A \subset [0,1]$). For $\varepsilon < 1/2$, set $A_\varepsilon = [\varepsilon, 1-\varepsilon]$. Then $d_{\rm H}(A_\varepsilon,[0,1]) = \varepsilon$ since the ε-neighborhood of A_ε is the smallest one which contains $[0,1]$. To compute $d_{\rm H}(F^n(A_\varepsilon), F^n([0,1])) = d_{\rm H}(F^n(A_\varepsilon),[0,1])$, we need to determine the smallest neighborhood of $F^n(A_\varepsilon)$ which contains $[0,1]$. Since $f_1([0,1]) = [0,1/2]$ and $f_2([0,1]) = [1/2,1]$, and f_1 and f_2 are both increasing, the leftmost point in $F^n(A_\varepsilon) = F^n([\varepsilon,1-\varepsilon])$ is $f_1^n(\varepsilon) = \varepsilon/(1+n\varepsilon)$. Therefore $d_{\rm H}(F^n(A_\varepsilon),[0,1]) \geq \varepsilon/(1+n\varepsilon)$, so

$$\frac{d_{\rm H}(F^n(A_\varepsilon), F^n([0,1]))}{d_{\rm H}(A_\varepsilon,[0,1])} \geq \frac{\varepsilon/(1+n\varepsilon)}{\varepsilon} = \frac{1}{1+n\varepsilon},$$

which can be made arbitrarily close to 1 using sufficiently small ε. (When $n = 1$ and $n = 2$ this lower bound on the ratio is actually an equality. I have not bothered to check if there is equality for all $n \geq 1$.)

Appendix A. More on the Hausdorff metric

In this appendix, we discuss two topics: how convexity behaves in the Hausdorff metric and how to isometrically embed $\mathrm{H}(X)$ into a space of continuous functions on X.

In a real vector space, a subset A is called *convex* when for any v and w in A the line segment $\{\lambda v + (1-\lambda)w : 0 \leq \lambda \leq 1\}$ between v and w lies in A. How do compact convex subsets of $\mathbf{R}^n$ behave in the Hausdorff metric?

Theorem A.1 (Blaschke)**.** *The compact convex subsets of $\mathbf{R}^n$ are a closed subset of $\mathrm{H}(\mathbf{R}^n)$: if $C_i \to C$ in $\mathrm{H}(\mathbf{R}^n)$ and each C_i is compact and convex then C is compact and convex.*

Proof. We already know the compact subsets of $\mathbf{R}^n$ are a closed subset of $\mathrm{H}(\mathbf{R}^n)$, so C is compact since each C_i is. Now suppose each C_i is convex and C is not convex. Pick $v, w \in C$ and $\lambda \in (0,1)$ such that $u := \lambda v + (1-\lambda)w \notin C$. Since the complement of C is open, there is an $\varepsilon > 0$ such that $B_\varepsilon(u) \cap C = \emptyset$.

For all large i, $d_{\rm H}(C_i, C) < \varepsilon/3$. Fix such an i. By (2.1), there are v_i and w_i in C_i such that

$$||v - v_i|| < \frac{\varepsilon}{3}, \quad ||w - w_i|| < \frac{\varepsilon}{3},$$

where $||\cdot||$ is the usual norm on $\mathbf{R}^n$. By convexity of C_i, the vector $u_i = \lambda v_i + (1-\lambda)w_i$ is in C_i, where we use the same λ as above. Then

$$\begin{aligned} ||u - u_i|| &= ||\lambda(v - v_i) + (1-\lambda)(w - w_i)|| \\ &\leq \lambda||v - v_i|| + (1-\lambda)||w - w_i|| \\ &< \frac{\varepsilon}{3}. \end{aligned}$$

Since $B_\varepsilon(u) \cap C = \emptyset$ and u_i is within $\varepsilon/3$ of u, any $x \in C$ satisfies $||x - u_i|| \geq 2\varepsilon/3$. Therefore $d_{\rm H}(C_i, C) \geq 2\varepsilon/3$, which is a contradiction. □

Corollary A.2. *The compact convex subsets of $\mathbf{R}^n$ are a locally compact metric space with respect to the Hausdorff metric. More precisely, for $R > 0$ the compact convex subsets of $\overline{B}_R(\mathbf{0}) \subset \mathbf{R}^n$ are a compact metric space: any sequence of compact convex subsets of $\overline{B}_R(\mathbf{0}) \subset \mathbf{R}^n$ has a convergent subsequence with respect to $d_{\rm H}$.*

Proof. Since $\overline{B}_R(\mathbf{0})$ is compact, $\mathrm{H}(\overline{B}_R(\mathbf{0}))$ is a compact subset of $\mathrm{H}(\mathbf{R}^n)$. By Theorem A.1, the set of compact convex subsets of $\overline{B}_R(\mathbf{0})$ is closed in $\mathrm{H}(\overline{B}_R(\mathbf{0}))$, and a closed subset of a compact space is compact. □

Corollary A.2 is called the Blaschke selection theorem. It was used by Blaschke in a proof of the isoperimetric theorem in $\mathbf{R}^n$: among all convex compact subsets of $\mathbf{R}^n$ with a fixed volume, the ones with the least surface area are spheres. The compactness of $\mathrm{H}(\overline{B}_R(\mathbf{0}))$ is used to prove the *existence* of a surface area minimizer, and extra work is needed to show it is a sphere. (That surface area minimizers should be convex follows from an argument using convex hulls.) For further details, see [7, pp. 104–105]. While Hutchinson's theorem and its application to fractals are a nice application of the completeness property of the Hausdorff metric, Blaschke's work on the isoperimetric theorem is a good application of the compactness properties of the Hausdorff metric. Blaschke's selection theorem is used to study many problems in convex geometry (see, for instance, [3, Sect. 2]). A compactness theorem of Mahler concerning the space of lattices in $\mathbf{R}^n$ can be proved with the Blaschke selection theorem by associating to each lattice a certain compact convex set (the Voronoi domain of the lattice) [5, pp. 412–419].

There is a notion of convexity in abstract metric spaces: call a metric space X convex if for any $x \neq y$ in X there is some $z \neq x, y$ in X such that $d(x,y) = d(x,z) + d(z,y)$. When X is a *closed* subset of $\mathbf{R}^n$ and is equipped with the *standard* metric coming from $\mathbf{R}^n$ then X is a convex metric space if and only if it is a convex subset of $\mathbf{R}^n$ in the classical sense of containing the line segment between any two points. (This equivalence of the abstract and concrete notions of convexity for subsets of $\mathbf{R}^n$ is not generally true when either italicized word above is relaxed. For instance, an open set in $\mathbf{R}^n$ need not be a convex subset of $\mathbf{R}^n$ but it is a convex metric space in the abstract sense when it is given the standard metric from $\mathbf{R}^n$. A sphere in $\mathbf{R}^n$ is also not a convex subset of $\mathbf{R}^n$ but it is a convex metric space using its surface metric.)

When $d(x,y) = d(x,z) + d(z,y)$, we say z lies *between* x and y. (In $\mathbf{R}^n$ with its standard metric, the points between two points are those on the line segment connecting them. On a sphere with its surface metric, all points besides the north and south poles lie between the poles.) Although an expression like $\lambda x + (1-\lambda)y$ for $\lambda \in [0,1]$ makes no sense in an abstract metric space, for any point z between x and y the ratio $d(x,z)/d(x,y)$ lies in $[0,1]$ and this number can be considered as a measure of how much closer z is to x than to y (a substitute for the number λ in classical convexity). Assuming X is convex and complete, all numbers in $[0,1]$ actually occur as such ratios:

Lemma A.3. *If X is a complete convex metric space then for any distinct points x and y in X and $\lambda \in [0,1]$ there is a $z \in X$ such that $d(x,y) = d(x,z)+d(z,y)$ and $d(x,z)/d(x,y) = \lambda$.*

Proof. See [4, Theorem 14.1, p. 41], which proves the stronger result that there is an isometric embedding of a real interval of length $d(x,y)$ into X which sends the two endpoints of the real interval to x and y. $\square$

Now we can generalize Theorem A.1 beyond $\mathbf{R}^n$.

Theorem A.4. *Let X be a complete convex metric space. The closed and bounded convex subsets of X form a closed subset of $\mathrm{H}(X)$: if $C_i \to C$ in $\mathrm{H}(X)$ and each C_i is convex then C is convex.*

Proof. Suppose C is not convex. Then there are $x \neq y$ in C such that for all $p \neq x, y$ in C

$$d(x,y) < d(x,p) + d(p,y).$$

Fix such a pair of points x and y. Pick $\lambda \in (0,1)$ and $\delta > 0$. Both will be specified later.

Since X is convex, there is a $z \neq x, y$ in X such that

$$d(x,y) = d(x,z) + d(z,y), \quad \frac{d(x,z)}{d(x,y)} = \lambda.$$

In particular, $z \notin C$ and z depends on λ. Since $X - C$ is open there is an open ball around z not meeting C, say $B_\varepsilon(z) \cap C = \emptyset$ for some $\varepsilon > 0$. (Now ε depends on λ too.)

For $i \gg 0$, $d_{\mathrm{H}}(C_i, C) \leq \delta$. Fix such an i (so i depends on δ). Then there are x_i and y_i in C_i such that $d(x_i, x) \leq \delta$ and $d(y_i, y) \leq \delta$. Using Lemma A.3, there is a $z_i \in C_i$ such that

$$d(x_i, y_i) = d(x_i, z_i) + d(z_i, y_i), \quad \frac{d(x_i, z_i)}{d(x_i, y_i)} = t.$$

Our goal is to show z_i is not too close to C, contradicting $d_{\mathrm{H}}(C_i, C) \leq \delta$. We have

$$\begin{aligned} d(z, z_i) &\leq d(z,x) + d(x, x_i) + d(x_i, z_i) \\ &= \lambda d(x,y) + d(x, x_i) + \lambda d(x_i, y_i) \\ &\leq \lambda(d(x,y) + d(x_i, y_i)) + \delta. \end{aligned}$$

Since $d(x_i, y_i) \leq d(x_i, x) + d(x,y) + d(y, y_i) \leq 2\delta + d(x,y)$,

$$d(z, z_i) \leq 2\lambda(d(x,y) + \delta) + \delta.$$

For any $x' \in C$, disjointness of C and $B_\varepsilon(z)$ gives $d(x', z) \geq \varepsilon$, so

$$\varepsilon \leq \varepsilon \leq d(x', z) \leq d(x', z_i) + d(z_i, z) \leq d(x', z_i) + 2\lambda(d(x,y) + \delta) + \delta.$$

Thus

$$d(x', z_i) \geq \varepsilon - 2\lambda(d(x,y) + \delta) - \delta.$$

This holds for all $x' \in C$, hence

$$d_{\mathrm{H}}(C, C_i) \geq \varepsilon - 2\lambda(d(x,y) + \delta) - \delta,$$

so $\varepsilon - 2\lambda(d(x,y) + \delta) - \delta \leq \delta$. Thus $\varepsilon \leq 2\lambda(d(x,y) + \delta) + 2\delta$, which is NOT a contradiction since $\varepsilon \leq \lambda d(x,y)$. □

Appendix B. Alternate Description of Hausdorff Metric

Now we describe an isometric embedding of $\mathrm{H}(X)$ into the space $C_b(X, \mathbf{R})$ of continuous bounded $\mathbf{R}$-valued functions on X. The set $C_b(X, \mathbf{R})$ is a real vector space metrized by the sup-norm metric: $\sup_{x \in X} |f(x) - g(x)|$. (The space $C(X, \mathbf{R})$ of *all* $\mathbf{R}$-valued continuous functions on X isn't metrized in this way unless all continuous functions are bounded, which need not happen: consider $X = \mathbf{R}$.) Here X is any metric space, not necessarily complete.

To send $\mathrm{H}(X)$ into $C_b(X, \mathbf{R})$ we use the distance functions δ_A introduced in (2.7): $\delta_A(x) = \operatorname{dist}(x, A)$. Previously we used these functions for compact A, but we no longer assume A is compact, so in particular we can't say that $\delta_A(x) = d(x, a)$ for some $a \in A$. From the proof of Lemma 2.27, $\delta_A \colon X \to \mathbf{R}$ is a continuous function.

Pick $A \in \mathrm{H}(X)$. We have $\delta_A(x) = 0$ if and only if x is a limit point of a sequence in A, which is the same as saying $x \in A$ since A is closed. Thus $A = \{x \in X : \delta_A(x) = 0\}$, so we can recover the set A from the function δ_A by seeing where the function is 0. The function $\delta_A \colon X \to \mathbf{R}$ need not be bounded: consider $X = \mathbf{R}^n$ with A a disc and $||x||$ large. However, the difference of two such functions is bounded:

Theorem B.1. *For A and B in $\mathrm{H}(X)$, the difference $\delta_A - \delta_B$ is a bounded function on X.*

This theorem will be superceded soon, but we include this intermediate result as practice with making estimates.

Proof. Pick $x \in X$ and $\varepsilon > 0$. Then there are $a_0 \in A$ and $b_0 \in B$ such that $d(x, a_0) < \delta_A(x) + \varepsilon$ and $d(x, b_0) < \delta_B(x) + \varepsilon$. Therefore

$$\delta_A(x) \leq d(x, a_0) < \delta_A(x) + \varepsilon \text{ and } \delta_B(x) \leq d(x, b_0) < \delta_B(x) + \varepsilon.$$

Now we can make estimates from above and below on $\delta_A(x) - \delta_B(x)$:

$$(d(x, a_0) - \varepsilon) - d(x, b_0) < \delta_A(x) - \delta_B(x) < d(x, a_0) - (d(x, b_0) - \varepsilon). \tag{B.1}$$

From the triangle inequality with the points x, a_0, and b_0, $-d(a_0, b_0) \leq d(x, a_0) - d(x, b_0) \leq d(a_0, b_0)$, so (B.1) implies

$$-d(a_0, b_0) - \varepsilon < \delta_A(x) - \delta_B(x) < d(a_0, b_0) + \varepsilon.$$

Hence

$$|\,\delta_A(x) - \delta_B(x)| \leq d(a_0, b_0) + \varepsilon \leq \sup_{a \in A, b \in B} d(a_0, b_0) + \varepsilon,$$

where the supremum is finite since A and B are bounded. The supremum does not depend on ε, so letting $\varepsilon \to 0^+$ gives

$$|\,\delta_A(x) - \delta_B(x)| \leq \sup_{a \in A, b \in B} d(a, b).$$

This upper bound is independent of $x \in X$. □

Lemma B.2. *For any A and B in* $\mathrm{H}(X)$*,*

$$\sup_{x \in X}(\delta_A(x) - \delta_B(x)) = \sup_{b \in B} \delta_A(b) = \inf\{r \geq 0 : B \subset E_r(A)\}.$$

Since $\delta_A(x) - \delta_B(x) = \delta_A(x)$ when $x \in B$, this lemma means that we can restrict the supremum of $\delta_A - \delta_B$ from X to B (the zero-set of δ_B) without affecting the value of the supremum.

Proof. Easily $\sup_{x \in X}(\delta_A(x) - \delta_B(x)) \geq \sup_{b \in B} \delta_A(b)$. We want the reverse inequality and we need to identify the supremum over B with the infimum mentioned in the lemma (which is how the Hausdorff distance will eventually get involved).

Pick $\varepsilon > 0$ and $x \in X$. There is some $b_\varepsilon \in B$ such that

$$d(x, b_\varepsilon) < \delta_B(x) + \varepsilon.$$

For all $a \in A$,

$$d(x, a) \leq d(x, b_\varepsilon) + d(b_\varepsilon, a) < \delta_B(x) + \varepsilon + d(b_\varepsilon, a),$$

so

$$\delta_A(x) \leq d(x, a) < \delta_B(x) + \varepsilon + d(b_\varepsilon, a).$$

Subtracting,

$$\delta_A(x) - \delta_B(x) - \varepsilon < d(b_\varepsilon, a).$$

Here a is arbitrary in A, so taking the infimum over all a gives

$$\delta_A(x) - \delta_B(x) - \varepsilon \leq \inf_{a \in A} d(b_\varepsilon, a) = \delta_A(b_\varepsilon) \leq \sup_{b \in B} \delta_A(b).$$

There is no ε-dependence in the supremum over B, so let $\varepsilon \to 0^+$:

$$\delta_A(x) - \delta_B(x) \leq \sup_{b \in B} \delta_A(b).$$

Now take the supremum over all $x \in X$ to get the first equation in the theorem.

It remains to show

$$\sup_{b\in B} \delta_A(b) = \inf\{r \geq 0 : B \subset E_r(A)\}.$$

In the proof of Lemma 2.27, we proved such a result (with the roles of A and B reversed) by using compactness. But we no longer assume A and B are compact, so we need a different argument.

If $B \subset E_r(A)$ then for all $b \in B$, $\delta_A(b) \leq r$. Therefore $\sup_{b\in B} \delta_A(b) \leq r$, so

$$\sup_{b\in B} \delta_A(b) \leq \inf\{r \geq 0 : B \subset E_r(A)\}.$$

To prove the reverse inequality, pick $\varepsilon > 0$. For any $b_0 \in B$, there is some $a_0 \in A$ such that

$$d(a_0, b_0) < \delta_A(b_0) + \varepsilon \leq \sup_{b\in B} \delta_A(b) + \varepsilon.$$

This shows $B \subset E_r(A)$ with $r = \sup_{b\in B} \delta_A(b) + \varepsilon$. Therefore

$$\inf\{r \geq 0 : B \subset E_r(A)\} \leq \sup_{b\in B} \delta_A(b) + \varepsilon.$$

Let $\varepsilon \to 0^+$. $\square$

Now we can refine Theorem B.1.

Theorem B.3. *For any A and B in $\mathrm{H}(X)$, $\sup_{x\in X} |\,\delta_A(x) - \delta_B(x)| = d_{\mathrm{H}}(A, B)$.*

Proof. From the proof of Lemma B.2, for any $x \in X$ we have

$$\delta_A(x) - \delta_B(x) \leq \inf\{r \geq 0 : B \subset E_r(A)\}.$$

Similarly,

$$\delta_B(x) - \delta_A(x) \leq \inf\{r \geq 0 : A \subset E_r(B)\},$$

so

$$|\,\delta_A(x) - \delta_B(x)| \leq \inf\{r \geq 0 : A \subset E_r(B) \text{ and } B \subset E_r(A)\} = d_{\mathrm{H}}(A, B).$$

Taking the supremum over all $x \in X$,

$$\sup_{x\in X} |\,\delta_A(x) - \delta_B(x)| \leq d_{\mathrm{H}}(A, B).$$

For the reverse inequality, since $A \subset X$

$$\sup_{x\in X} |\,\delta_A(x) - \delta_B(x)| \geq \sup_{a\in A} \delta_B(a) = \inf\{r \geq 0 : A \subset E_r(B)\},$$

where the equality comes from Lemma B.2 (with the roles of A and B reversed). Similarly,

$$\sup_{x\in X} |\,\delta_A(x) - \delta_B(x)| \geq \sup_{b\in B} \delta_A(b) = \inf\{r \geq 0 : B \subset E_r(A)\}.$$

Therefore each of A and B is in the r'-expansion of the other, so

$$\sup_{x\in X} |\,\delta_A(x) - \delta_B(x)| \geq \max(\inf\{r \geq 0 : A \subset E_r(B)\}, \inf\{r \geq 0 : B \subset E_r(A)\}).$$

We want to show this maximum is at least as large as $d_{\mathrm{H}}(A, B)$. Pick $\varepsilon > 0$. There is $s < \inf\{r \geq 0 : A \subset E_r(B)\} + \varepsilon$ such that $A \subset E_s(B)$ and $t < \inf\{r \geq 0 : B \subset E_r(A)\} + \varepsilon$ such that $B \subset E_t(A)$. Let $r' = \max(s, t)$, so $A \subset E_{r'}(B)$ and $B \subset E_{r'}(A)$. Therefore

$$\begin{aligned} d_{\mathrm{H}}(A, B) &= \inf\{r \geq 0 : A \subset E_r(B) \text{ and } B \subset E_r(A)\} \\ &\leq r' \\ &< \max(\inf\{r \geq 0 : A \subset E_r(B)\} + \varepsilon, \inf\{r \geq 0 : B \subset E_r(A)\} + \varepsilon) \\ &= \max(\inf\{r \geq 0 : A \subset E_r(B)\}, \inf\{r \geq 0 : B \subset E_r(A)\}) + \varepsilon. \end{aligned}$$

Now let $\varepsilon \to 0^+$ to see that

$$d_{\mathrm{H}}(A, B) \leq \max(\inf\{r \geq 0 : A \subset E_r(B)\}, \inf\{r \geq 0 : B \subset E_r(A)\}),$$

so $\sup_{x \in X} |\delta_A(x) - \delta_B(x)| \geq d_{\mathrm{H}}(A, B)$. □

If X is compact then each $\delta_A \colon X \to \mathbf{R}$ is bounded since continuous functions on a compact set are bounded. Then Theorem B.3 says the mapping $\mathrm{H}(X) \to C_b(X, \mathbf{R}) = C(X, \mathbf{R})$ given by $A \mapsto \delta_A$ is an isometric embedding. If X is not compact, the functions δ_A have no reason to be bounded. But if we simply fix some $D \in \mathrm{H}(X)$ – even just a one-element set $D = \{d_0\}$ would do – and associate to each $A \in \mathrm{H}(X)$ the function $\delta_A - \delta_D$, which is bounded, then we get an isometric embedding of $\mathrm{H}(X)$ into $C_b(X, \mathbf{R})$. Indeed, the difference$(\delta_A - \delta_D) - (\delta_B - \delta_D) = \delta_A - \delta_B$ has sup-norm over X equal to $d_{\mathrm{H}}(A, B)$ by Theorem B.3. Thus, subject to the arbitrariness of choosing some (nonempty) closed and bounded subset $D \subset X$ as an anchor, we obtain an isometric embedding of $\mathrm{H}(X)$ into $C_b(X, \mathbf{R})$ by $A \mapsto \delta_A - \delta_D$. Here X is any metric space, complete or not. This isometric embedding justifies the "naturalness" of the Hausdorff metric on $\mathrm{H}(X)$, since it agrees with a sup-norm metric on functions.

In the course of proving Theorem B.3, we obtained a formula for the Hausdorff distance:

$$d_{\mathrm{H}}(A, B) = \max(\sup_{a \in A} \delta_B(a), \sup_{b \in B} \delta_A(b)) = \max(\sup_{a \in A} \operatorname{dist}(a, B), \sup_{b \in B} \operatorname{dist}(b, A)).$$

This formula is less intuitive than our original definition of the Hausdorff distance. What does it say? Glossing over the distinction between supremum and maximum, think of $\sup_{a \in A} \operatorname{dist}(a, B)$ as the farthest that a point in A can be from its nearest point in B, and $\sup_{b \in B} \operatorname{dist}(b, A)$ as the farthest that a point in B can be from its nearest point in A. The Hausdorff distance is the maximum of these two numbers.

References

[1] M. Barnsley, "Fractals Everywhere," Academic Press, Boston, 1988.

[2] D. G. Bennett and B. Fisher, *On a Fixed Point Theorem for Compact Metric Spaces*, Math. Magazine **47** (1974), 40–41.

[3] Y. Benyamini, *Applications of the Universal Surjectivity of the Cantor Set*, Amer. Math. Monthly **105** (1998), 832–839.

[4] L. M. Blumenthal, "Theory and Applications of Distance Geometry," Oxford Univ. Press, Oxford, 1953.

[5] W. A. Coppel, "Number theory: an introduction to mathematics. Part B," Springer-Verlag, New York, 2006.

[6] M. Edelstein, *On Fixed and Periodic Points Under Contractive Mappings*, J. London Math. Soc. **37** (1962), 74–79.

[7] H. G. Eggleston, "Convexity," Cambridge Univ. Press, Cambridge, 1958.

[8] J. Henrickson, *Completeness and Total Boundedness of the Hausdorff Metric*, MIT Undergraduate Journal of Mathematics **1** (1999), 69–79. Also available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.353.633.

[9] S. G. Hoggar, “Mathematics for Computer Graphics,” Cambridge Univ. Press, Cambridge, 1993.
[10] J. E. Hutchinson, *Fractals and Self-Similarity*, Indiana Univ. Math. J. **30** (1981), 713–747.
[11] H. Katsuura, *Continuous Nowhere-Differentiable Functions – an Application of Contraction Mappings*, Amer. Math. Monthly **98** (1991), 411–416.