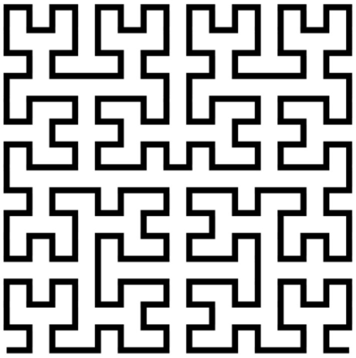
Space-filling Curves



Kazimierz Wilowski
Supervisor: Professor Lars Olsen
8 April 2022

Abstract

A space-filling curve is a curve which travels through every point in an area or volume (for example, the unit square). Space-filling curves are interesting examples of objects which can challenge the mathematical intuitions we hold with regard to certain concepts like dimension and volume. This project gives an overview of the topic of space-filling curves, including a discussion of the historical context and background which led to interest in space-filling curves within the mathematical community, some key mathematical results necessary for understanding space-filling curves, and an extended discussion and investigation of some of the properties of a specific example of a space-filling curve.

Declaration

I certify that this project report has been written by me, is a record of work carried out by me, and is essentially different from work undertaken for any other purpose or assessment.

Contents

1 Introduction

Towards the end of the 19th century, Georg Cantor proved the existence of a variety of counter-intuitive one-to-one correspondences between different sets, including between the unit interval $[0, 1]$ and the n-dimensional unit cube $[0, 1]^n$ [17]. Throughout his mathematical career Cantor endured regular public opposition and polemical criticism aimed at many of his more counter-intuitive ideas and theories, and his proofs for the existence of these one-to-one correspondences were no exception - fellow German mathematician Leopold Kronecker unsuccessfully campaigned Cantor to withdraw one of his papers elucidating on such ideas [5]. The notion that there could be a bijection from $[0, 1]$ to $[0, 1]^n$ was indeed a shocking result since it cast doubt on how sturdy the seemingly intuitive mathematical conception of dimension really was. Dimension had up to then been taken as a relatively intuitive and straightforward concept and the fundamental distinction between different dimensions was considered as self-evident. Were there to exist a continuous bijection from $[0, 1]$ to $[0, 1]^2$ (or $[0, 1]^3$, etc.) it would raise serious questions about what it truly meant for two dimensions to be distinct from one and other and indicate that the intuitive notion of dimension was ill defined, and that developing a rigorous, sensible notion of dimension might prove challenging [10]. Such questions subverting what had until then been taken as intuitive, self-evident fact seem to foreshadow the intense debates and discussions regarding the foundations of mathematics which would come to characterise a great deal of mathematics in the latter part of the 19th and the first part of the 20th century. In 1879 Eugen Netto would restore faith in the intuitive notion of dimension by proving any such bijection from $[0, 1]$ to $[0, 1]^2$ (or $[0, 1]^3$, etc.) would necessarily be discontinuous [19][16], indicating that the straight-forward intuitive notion of dimension did in fact provide a mathematically sound basis for establishing a mathematically rigorous notion of spatial dimension.

If it was impossible to construct a continuous bijection from $[0, 1]$ to $[0, 1]^2$, a natural question was then to ask whether there was a continuous, surjective function from $[0, 1]$ to $[0, 1]^2$. The answer to this question was yes, and the first of these so-called *"space-filling curves"* was constructed in 1890 by Giuseppe Peano, and the existence of the eponymous Peano curve provided another entry into the expanding list of counter-intuitive results emerging during the time [18]. Perhaps surprisingly, the article in which Peano described his curve contained no illustrations, and it wasn't until the following year, when David Hilbert described his own space-filling curve, accompanied with illustrations, that the recursive visual elegance which space-filling curves often possess was shown in a way appreciable to the layperson [11]. It is this family of continuous, surjective functions $f : [0, 1] \to [0, 1]^2$ (or more generally $f : [0, 1] \to K$ where $K \subseteq \mathbb{R}^n$ for $n > 1$ and K is a set which has non-zero area, volume, etc.[1]) which we call space-filling curves and which occupy the attention of this report.

We begin our discussion of the topic by investigating the previously mentioned key result of Netto, which states there can be no continuous, bijective mapping from the unit interval to the unit square. We then continue by engaging in an in depth discussion of the properties of a specific instance of a space-filling curve: the aforementioned Hilbert curve, in order to more intuitively grasp the nature and some of the properties which space-filling curves can possess. This discussion includes defining and verifying the space-filling properties of the Hilbert curve, as well as investigating a variety of interesting properties of the curve, such as its locality preserving nature, and the self-similar character of the curve's coordinate functions. Finally we conclude with an outlook section, briefly looking forward, towards a selection of more advanced ideas and concepts related to but beyond the scope of the material covered in this project and report.

[1] For an example of a space-filling curve with a more unusually shaped image, see the Gosper curve [9]

2 Netto's Theorem

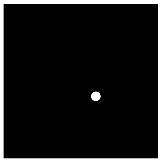
2.1 Introduction to Netto's Theorem

Netto's theorem is perhaps one of the most important results to consider when regarding space-filling curves, since it entails one of their key characteristics, namely that any space-filling curve necessarily maps multiple points from the unit interval to the same point in its image, i.e. it is impossible to have an injective space-filling curve.

Theorem 1 (Netto)**.** *Any bijective map* $f : [0,1] \longleftrightarrow [0,1]^2$ *is discontinuous.*

Conceptually, this result makes a certain amount of intuitive sense, since if we consider the interval $[0,1]$ with a single point removed, we see that it is necessarily split into two separate, disconnected sections, like cutting a piece of string:

However, for the unit square $[0,1]^2$, removing a single point still preserves the intuitive "wholeness" of the section:



This is similar to punching a hole in a piece of paper. The paper remains a single item, despite the hole, while the string becomes two discrete parts. It seems reasonable to conjecture that Netto's theorem holds true: Bijective continuous mapping should exist between two sets differing in this fundamental sense. Before proving Netto's theorem, we first lay some groundwork. In outlining a proof of Netto's theorem, we follow a similar route to that taken in [19] (pg. 85-95).

2.2 Definitions and Preliminary Results

Before we tackle the proof of Netto's theorem, we will first verify a handful of preliminary results and introduce and define some of the concepts which we will use in this section and further throughout our discussion of space-filling curves. Here and throughout, we notate the Euclidean distance between two points x and y in $\mathbb{R}^n$ as $|x-y|$.

Definition 1 (open balls)**.** *The open ball or neighbourhood at* x *with radius* r*, denoted* $B(x,r)$ *is the set* $\{y \,:\, |x-y| < r\}$.

Definition 2 (open sets)**.** *A set* $A \in \mathbb{R}^n$ *is open if for all* $a \in A$ *we can find* $r > 0$ *such that* $B(a,r) \subset A$.

Definition 3 (closed sets)**.** *A set* $A \in \mathbb{R}^n$ *is closed if its complement* A^c *is open.*

We will also make use of the following alternative characterisation of closed sets.

Lemma 1. *A set* $A \in \mathbb{R}^n$ *is closed if and only if it contains all points* $a \in \mathbb{R}^n$ *such that for any* $\delta > 0$, $B(a,\delta) \cap A \neq \emptyset$ *(such points are called accumulation points of* A*).*

Proof.
($\Rightarrow$) Suppose A is closed and a is an accumulation point of A but $a \notin A$, then $A \in A^c$, since A^c is open, there is $\delta > 0$ such that $B(a,\delta) \in A^c$ so $B(a,\delta) \cap A = \emptyset$, but this contradicts the original assumption that a was an accumulation point of A.

($\Leftarrow$) Assume A contains all of its accumutlation points, if $x \notin A$, there is a $\delta > 0$ such that $B(x,\delta)$ contains no points from A, implying that $B(x,\delta) \in A^c$. Since this is true for all x not in A, then A^c must be open, and so A is by definition closed. □

With our first few definitions stated, we can proceed to prove the following result which introduces a characterisation of function continuity in terms of open sets.

Lemma 2. *$f : K \longrightarrow \mathbb{R}^n$ where $K \subseteq \mathbb{R}^m$ is continuous if and only if for every open set $\Omega \subseteq \mathbb{R}^n$ there is an open set $\Omega_1 \subseteq \mathbb{R}^m$ such that $f^{-1}(\Omega) = \Omega_1 \cap K$*

Proof.
($\Rightarrow$) Let $\Omega \subseteq \mathbb{R}^n$ be an open set, if $f^{-1}(\Omega) = \emptyset = \emptyset \cap K$ then we are immediately done (since $\emptyset$ is trivially open). So assume $f^{-1}(\Omega) \neq \emptyset$. For each $f(x) \in \Omega$, since Ω is open there is some neighbourhood $B(f(x), \varepsilon) \subseteq \Omega$. Since f is continuous there is a neighbourhood around $x \in K$ such that $f(B(x, \delta)) \subset B(f(x), \varepsilon)$. We construct Ω_1 as:

$$\Omega_1 = \bigcup_{x \in f^{-1}(\Omega)} B(x, \delta)$$

Finally we prove that the two sets $\Omega_1 \cap K$ and $f^{-1}(\Omega)$ are equal. Ω_1 is open by construction, and if $x \in f^{-1}(\Omega), x \in \Omega_1$, so $f^{-1}(\Omega) \subseteq \Omega_1 \subseteq K \Rightarrow f^{-1}(\Omega) \subseteq \Omega_1 \cap K$.

If $x \in \Omega_1 \cap K$ then $x \in B(y, \delta)$ for some $y \in f^{-1}(\Omega)$, and so $f(x) \in \Omega$, so $x \in f^{-1}(\Omega) \Rightarrow (\Omega_1 \cap K) \subseteq f^{-1}(\Omega)$ so $\Omega_1 \cap K = f^{-1}(\Omega)$.

($\Leftarrow$) Let $a \in K$, for every $\varepsilon > 0$ the set $B(f(a), \varepsilon) = \Omega$ is open in $\mathbb{R}^n$ and $a \in f^{-1}(B(f(a), \varepsilon))$. By our hypothesis, there is an open set $\Omega_1 \in K$ such that $f^{-1}(B(f(a), \varepsilon)) = \Omega_1 \cap K$. Since $a \in \Omega_1$ and since Ω_1 is open, there is a $\delta > 0$ such that $B(a, \delta) \subseteq \Omega_1$. Since $B(a, \delta) \cap K \subseteq \Omega_1 \cap K = f^{-1}(B(f(a), \varepsilon))$, the condition for continuity is met. □

As we might imagine, the same characterisation of continuous functions also holds for closed sets as well.

Corollary to Lemma 2 *$f : K \longrightarrow \mathbb{R}^n$ is continuous if and only if for every closed set $\Omega \subseteq \mathbb{R}^n$ there is a closed set $\Omega_1 \subseteq \mathbb{R}^m$ such that $f^{-1}(\Omega) = \Omega_1 \cap K$*

Proof.
($\Rightarrow$) Let $\Omega \in \mathbb{R}^n$ be a closed set. Then Ω^c is open and so there exists an open set Ω_1^c such that $f^{-1}(\Omega^c) = \Omega_1^c \cap K$. If x is in $\Omega_1 \cap K$ then x cannot also be an element of Ω_1^c and so $f(x) \notin \Omega^c$, the only other option is that $f(x) \in \Omega$, so every element of $\Omega_1 \cap K$ must be an element of $f^{-1}(\Omega)$, i.e. $\Omega_1 \cap K \subseteq f^{-1}(\Omega)$.

If $x \in f^{-1}(\Omega)$ then $x \notin f^{-1}(\Omega^c)$ so $x \notin \Omega_1^c \cap K$ but x must be in K, so $x \notin \Omega_1^c$ implying $x \in \Omega_1$ so therefore $x \in \Omega_1 \cap K$ and so every element in $f^{-1}(\Omega)$ must also be an element of $\Omega_1 \cap K$, i.e. $f^{-1}(\Omega) \subseteq \Omega_1 \cap K$. So $f^{-1}(\Omega) = \Omega_1 \cap K$.

($\Leftarrow$) Let $a \in K$, for every $\varepsilon > 0$, the set $B(f(a), \varepsilon) = \Omega^c$ is open and so Ω is a closed set in $\mathbb{R}^n$. By our hypothesis, there is a closed set $\Omega_1 \in K$ such that $f^{-1}(\Omega) = \Omega_1 \cap K$. We note that $f^{-1}(B(f(a), \varepsilon)) \nsubseteq \Omega_1 \cap K$. Therefore $f^{-1}(B(f(a), \varepsilon)) \subseteq (\Omega_1)^c \cap K = \Omega_1^c \cap K$. Since Ω_1^c is the complement of a closed set it is open, since $f(a) \in \Omega^c$, it must be the case that $a \in \Omega_1^c$ and so there is some neighbourhood $B(a, \delta) \subseteq \Omega_1^c$. Finally $f(B(a, \delta)) \subseteq f(\Omega_1^c) = \Omega^c = B(f(a), \varepsilon)$ so $f(B(a, \delta)) \subseteq B(f(a), \varepsilon)$ and the condition for continuity is met. □

An important concept used in our proof of Netto's theorem is the notion of *compactness*, which we introduce here.

Definition 4 (Compactness)**.** *A set T is said to be compact if and only if every open cover of T contains a finite subcover.*

Next we prove a helpful lemma which states that continuous functions map compact sets to compact sets.

Lemma 3. *For a continuous function $f : K \longrightarrow \mathbb{R}^n$, where $K \subseteq \mathbb{R}^m$ and $T \subseteq K$ is compact, $f(T)$ will also be a compact set.*

Proof. Let $\{G_\alpha | \alpha \in A\}$ be an open cover of $f(T)$. It must be the case that $\{f^{-1}(G_\alpha) | \alpha \in A\}$ is an open cover of T, since for all $x \in T, f(x) \in G_\alpha$ for some α, so $x \in f^{-1}(G_\alpha)$. Since T is compact, $\{f^{-1}(G_\alpha) | \alpha \in A\}$ must contain a finite subcovering of T: $\{f^{-1}(G_1), f^{-1}(G_2), f^{-1}(G_3), ...f^{-1}(G_j)\}$. We now observe that for each $f(x) \in f(T)$, $x \in f^{-1}(G_i)$ for some $i \in \{1, 2, 3, ...j\}$ and so $f(x) \in G_i$, thus $\{G_1, G_2, G_3, ...G_j\}$ is a finite cover covering $f(T)$, so our original cover of $f(T)$ contains a finite subcover and so $f(T)$ is compact. □

Another result regarding compact sets states that any closed subset of a compact set is itself compact.

Lemma 4. *For a compact set $K \subseteq \mathbb{R}^n$, if $T \subseteq K$ and T is closed, then T is also compact.*

Proof. Let C_T be an open cover for T, $T^c = \mathbb{R}^n \backslash T$ is an open set and so $C_K = C_T \cup \{T^c\}$ is an open cover of $\mathbb{R}^n$ and hence an open cover of K. C_K must then contain a finite subcover of K, C'_K. Since $T^c \cup T = \emptyset$, $C'_K \backslash \{T^c\}$ must also be a finite subcover for T, but $C'_K \backslash T^c \subseteq C_T$, so the original open cover contains a finite subcover and so T is compact. □

We now proceed in proving a helpful theorem. The Heine-Borel Theorem provides a significantly more convenient characterisation of compactness which will allow us to very easily show that a variety of sets are compact with very little work.

Theorem 2 (Heine-Borel)**.** *A set $K \in \mathbb{R}^n$ is compact if and only if it closed and bounded.*

Proof.
($\Rightarrow$)
K compact $\Rightarrow$ K closed: Assume K does not contain one of its accumulation points, a. We construct the following cover for K: $\{\Omega_j | j = 1, 2, 3...\}$ where $\Omega_j = \{x \in \mathbb{R}^n | \; |x - a| > \frac{1}{j}\}$. Since K is compact, this cover contains a finite subcover $\{\Omega_j | j = j_1, j_2, j_3, ..., j_n\}$, let $j_0 = \max\{j_1, j_2, j_3, ...j_n\}$, then $K \subseteq \Omega_{j_0}$. This means that $\{x \in \mathbb{R}^n | \; |x - a| \leq \frac{1}{j_0}\}$ contains no points in K, therefore a cannot be an accumulation point of a and so the original assumption that K does not contain one of its accumulation points must be false and so K is closed.

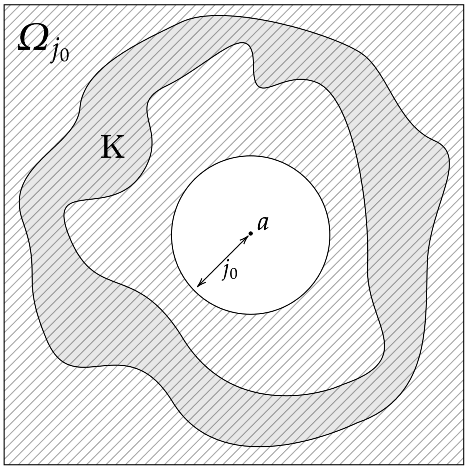


Figure 1: The set Ω_{j_0} contains all of K but does not contain and leaves space around a

K compact $\Rightarrow$ K bounded: The set $\{B(0, k) | k = 1, 2, 3, ...\}$ is clearly an open cover of K. Since K is compact, there is a finite subcover $\{B(0, k_1), B(0, k_2), B(0, k_3), ...B(0, k_j)\}$. Let $k_0 = \max(k_1, k_2, k_3, ...k_j)$, then $K \subseteq \bigcup_{i=1}^{j} B(0, k_i)(0) \subseteq B(0, k_0) = \{x \in \mathbb{R}^n | \; |x| < k_0\}$. Clearly $B(0, k_0)$ is bounded therefore K is too.

($\Leftarrow$)
Since K is closed and bounded, it can be contained in a closed n-dimensional cube W. Since K is closed, by **Lemma 4** it is sufficient to show that W is compact to show K is compact. We assume W is not compact, so does not have the Heine-Borel property. Let $\Omega = \{\Omega_\alpha | \alpha \in A\}$ be an open cover of W containing no finite subcover. We partition W into 2^n subcubes each with side length $\frac{1}{2}$. In order for Ω not to have a finite subcover for the whole of W, at least one subcube, say W_1, must also not be covered by any finite subset of Ω (otherwise, if all subcubes had finite covers from Ω, we could simply take the union of the subcubes' finite subcovers to find a subcover for the whole of W). We now split W_1 into 2^n subcubes and again take one of the subcubes without a finite cover in Ω, and label it W_2. We then repeat this process to yield a series of nested, closed subcubes $W_1 \supset W_2 \supset W_3 \supset ...$, note W_k

has side length $\frac{1}{2^k}$. Taking an arbitrary point x_k from each set W_k yields a Cauchy sequence which in the complete metric space of $\mathbb{R}^n$ with the Euclidean norm must converge to a unique point $p \in W$, and since each W_k is closed, $p \in W_k$ for all k. It must be the case that $p \in \Omega_{\alpha_0}$ for some $\alpha_0 \in A$. Since Ω_{α_0} is open, there must be some δ such that $B(p, \delta) \subseteq \Omega_{\alpha_0}$, but also for sufficiently large k, $W_k \subseteq B(p, \delta) \subseteq \Omega_{\alpha_0}$, meaning W_k can be covered using a single element of Ω, contradictory to the assertion that none of the subcubes in the sequence could be covered by finite subsets of Ω. Therefore $[0, 1]^n$ must be compact. □

We conclude this section by proving a couple of final helpful lemmas which we will make use of later in the process of proving Netto's theorem. Firstly we show that the inverse of a continuous, injective function from a compact set must also be continuous.

Lemma 5. *If a function $f : K \longrightarrow \mathbb{R}^n$, where K is compact and $K \subseteq \mathbb{R}^m$, is continuous and injective, then its inverse $f^{-1} : f(K) \longrightarrow K$ is also continuous.*

Proof. Let $A \subseteq \mathbb{R}^m$ be a closed set, since K is compact and therefore closed $A \cap K \subseteq K$ is closed and so is a closed subset of a compact set and is therefore by **Lemma 4** is itself compact. By **Lemma 3** and **Theorem 2**, $f(A \cap K) = A_1$ is compact and so is also closed. Let $g = f^{-1}$, then $f(A \cap K) = A_1 = g^{-1}(A)$. Since $A_1 \subseteq f(K) = g^{-1}(K)$, we can write $g^{-1}(A) = f(A \cap K) \cap g^{-1}(K) = A_1 \cap g^{-1}(K)$, and finally since A_1 is closed, by **Corollary to Lemma 2**, $g = f^{-1}$ is continuous. □

Finally we introduce the notion of pathwise connectedness and prove a basic result stating that continuous functions preserve pathwise connectedness.

Definition 5. *A set A is pathwise connected if for all $a, b \in A$, there is a continuous function $\gamma : [0, 1] \to A$ such that $\gamma(0) = a$ and $\gamma(1) = b$.*

Lemma 6. *If $f : K \longrightarrow \mathbb{R}^n$ is a continuous function and $T \subseteq K$ is a pathwise connected subset of K, then $f(T)$ is also pathwise connected.*

Proof. Let $y_1, y_2 \in f(T)$, there exists $x_1, x_2 \in T$ such that $f(x_1) = y_1$ and $f(x_2) = y_2$. Since T is pathwise connected, there exists a continuous function $\gamma : [0, 1] \longrightarrow T$ such that $\gamma(0) = x_1$ and $\gamma(1) = x_2$. Let $\Gamma : [0, 1] \longrightarrow f(T)$ be defined by $\Gamma(x) = (f \circ \gamma)(x)$. Since f and γ are both continuous, so is Γ. $\Gamma(0) = (f \circ \gamma)(0) = f(\gamma(0)) = f(x_1) = y_1$ and $\Gamma(0) = (f \circ \gamma)(1) = f(\gamma(1)) = f(x_2) = y_2$, so $f(T)$ is also pathwise connected. □

2.3 Proofs of general results concerning $[0, 1]$ and $[0, 1]^2$

We have now introduced most of the machinery required to prove Netto's theorem. Before proceeding to the final statement of the proof however, we must verify a few different properties of the unit interval and unit square, so as to be sure that the results explored in the prior section can be applied as required to prove Netto's theorem.

Lemma 7. *$[0, 1]^2$ is compact.*

Proof. As part of the proof of the Heine-Borel Theorem we saw that any n-dimensional cube W is compact. $[0, 1]^2$ is simply the specific case of this, so is compact. □

Lemma 8. *$[0, 1]^2 \backslash t_0$ is pathwise connected for all points t_0.*

Proof. Given two points p_1, p_2 such that $p_1 \neq p_2$, if t_0 is not colinear with p_1 and p_2 we can construct a continuous straight path between p_1 and p_2 so we are done. Otherwise, take a third point, p_3 not colinear with the other three points. Since it is not colinear with the previous points, the straight lines from p_1 to p_3 and from p_2 to p_3 do not go through t_0, therefore we have a continuous path from p_1 to p_2 by first moving from p_1 to p_3 then to p_2. □

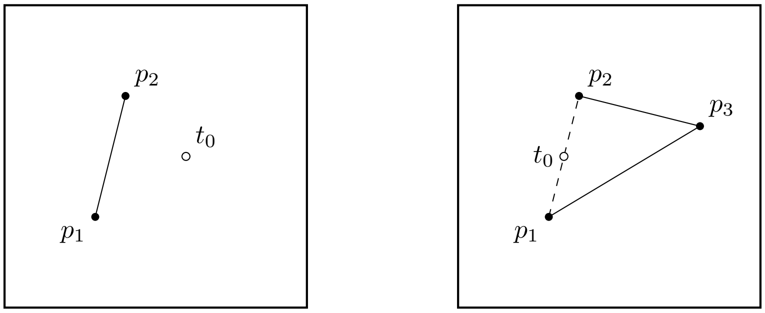


Figure 2: The two different cases arising when considering whether a continuous path exists between arbitrary points in $[0,1]^2\backslash t_0$. Right: t_0, p_0 and p_1 not colinear. Left: t_0, p_0 and p_1 colinear.

Lemma 9. *Let $t_0 \in (0,1)$, then $[0,1]\backslash t_0$ is not pathwise connected.*

Proof. Assume $[0,1]\backslash t_0$ is pathwise connected, then there exists a continuous function, $\gamma : [0,1] \to [0,1]\backslash t_0$ such that $\gamma(0) = 0$ and $\gamma(1) = 1$. By the intermediate value theorem, γ must take every value $y \in [0,1]$, however γ never takes the value t_0 since clearly t_0 is not contained in $\gamma([0,1]) = [0,1]\backslash t_0$. Therefore no such γ can exist, and so $[0,1]\backslash t_0$ cannot be pathwise connected. □

2.4 Proof of Netto's Theorem

Finally, we conclude this first chapter of the report by bringing together the results explored so far to give a proof for Netto's theorem. The variety of results and lemmas we have already investigated means that the actual final proof for Netto's theorem given here is relatively short.

Proof. We will proceed by assuming there exists $f : [0,1] \longrightarrow [0,1]^2$ such that f is a bijective and continuous function and derive a contradiction. f is continuous and injective, so by **Lemma 5** f^{-1} is also continuous.

Since f^{-1} is continuous, by **Lemma 6**, then it will map pathwise connected sets to pathwise connected sets. We remove a point $t_0 \in (0,1)$ from $[0,1]$ and its image $f(t_0)$ from $[0,1]^2$ and observe that $[0,1]^2\backslash f(t_0)$ is pathwise connected but $[0,1]\backslash t_0$ is not (**Lemma 8** and **9**) and so the continuous function f^-1 seems to map a pathwise connected set to a set which connected, therefore we have reached a contradiction and the original assumption that f was continuous must be false. □

We conclude this section by noting that the proof of Netto's theorem given here can be relatively straight-forwardly generalised to mappings from the unit interval to higher dimensional boxes ($[0,1]^3$, $[0,1]^4$, etc.), since all these intervals are also closed and bounded and possess the same property that removing a single point from them maintains their pathwise connectedness. In fact, the theorem can be further generalised to any higher dimensional set which also possesses these properties, which includes the image of any space-filling curve: since by **Lemma 3**, the image of any space-filling curve $f([0,1])$, must be compact (closed and bounded) since it is the image of a continuous function of the unit interval which is a compact set, and since the image of a space-filling curve will by definition have non-zero n-dimensional area, it will be straightforward to remove a point $f(t_0)$ for $t \in (0,1)$ from the area of the curve's image such that $f([0,1])\backslash f(t_0)$ remains pathwise-connected but $[0,1]\backslash t_0$ does not.

3 A Concrete Example: The Hilbert Curve

Beginning in this section, and for the rest of this report, we will focus our attention on a specific example of a space-filling curve, and the properties this example possesses. The space-filling curve we will investigate is the Hilbert curve - perhaps the most well known space-filling curve - which was first described by David Hilbert in 1891 [11]. The Hilbert curve is a function $H : [0,1] \to [0,1]^2$ which is continuous and surjective but not injective.

3.1 Constructing the Hilbert approximation curves

The Hilbert curve is perhaps most easily conceptualised as the uniform limit of an infinite sequence of approximating curves $(H_n)_{n\in\mathbb{N}}$ which can be constructed recursively. First we define the first Hilbert approximation curve $H_0 : [0,1] \to [0,1]^2$, which all subsequent approximation curves will be recursively defined in terms of. H_0 follows the continuous path beginning at $(\frac{1}{4}, \frac{1}{4})$ travels to $(\frac{1}{4}, \frac{3}{4})$ then $(\frac{3}{4}, \frac{3}{4})$ then $(\frac{1}{4}, \frac{1}{4})$. These points being the centers of the four quadrants of the unit square.

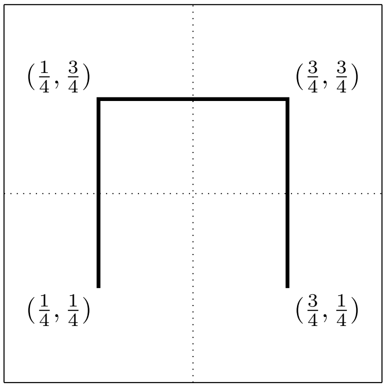

$(\frac{1}{4}, \frac{3}{4})$
$(\frac{3}{4}, \frac{3}{4})$
$(\frac{1}{4}, \frac{1}{4})$
$(\frac{3}{4}, \frac{1}{4})$

Figure 3: The first Hilbert approximation curve H_0

To generate further approximate curves, a recursive algorithm is used: To generate H_{n+1} we take four copies of H_n and scale them by a factor of a half. Then each of the four copies become one of the four quadrants of the new curve iteration. The copies used for the bottom left and right quadrants are rotated 90° clockwise and 90° anti-clockwise respectively, and the end of the curves in each quadrant are joined to the start of the curve in the next quadrant, thus producing the next iteration.

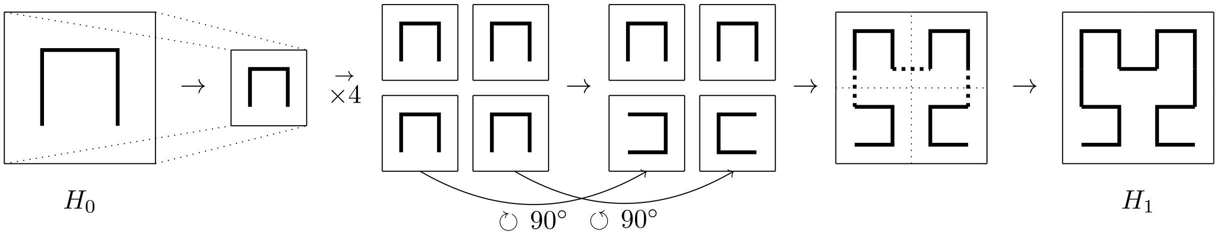


Figure 4: Illustration of the process of generating H_1 from H_0

The process outlined above can be used to generate more approximation curves, e.g. H_2 is yielded when the process is performed starting with H_1 rather than H_0. In general, H_{n+1} can be generated using H_n. Below are the first 4 hilbert approximation curves.

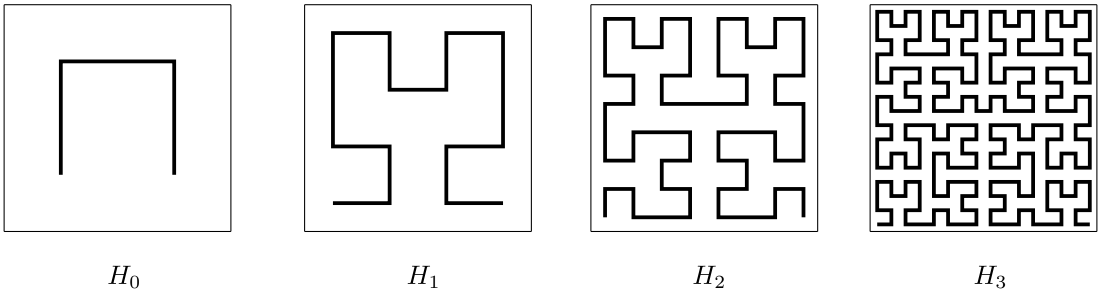


Figure 5: The first four Hilbert approximation curves

The Hilbert curve is taken to be the function which is the uniform limit tended to with respect to $||\cdot||_\infty$ by this sequence of functions: $H = \lim\limits_{n\to\infty} H_n$. It quickly becomes inconvenient and unwieldy to write explicit piecewise formulas for higher order Hilbert approximation curves, however it is still possible to work with and explore the properties of the Hilbert curve and its approximations using the recursive, geometric perspective outlined in this section. Below we will show the uniform limit of the sequence $(H_n)_{n\in\mathbb{N}}$ with respect to $||\cdot||_\infty$, which we define to be the Hilbert curve, exists.

3.2 Definition of the Hilbert space-filling curve

In this section we more rigorously define the Hilbert curve using precise mathematical language. To do this we will show that the sequence $(H_n)_{n\in\mathbb{N}}$ converges uniformly in the metric space of continuous functions from $[0,1]$ to $[0,1]^2$ with the infinity norm. It is the limit that this sequence tends to uniformly with respect to the infinity norm which we define to be the Hilbert curve $H : [0,1] \to [0,1]^2$.

Lemma 10. $||H_n - H_{n+1}||_\infty \leq \frac{\sqrt{2}}{2^n}$.

Proof. Up to now the rate of traversal of the Hilbert curves has not been mentioned, and up to now it has been of no consequence. We take the rate of traversal of the curve H_n to be such that when the unit square is split into a grid of 4^n equally sized sub-squares, the pre-image of each sub-square is an interval of diameter $\frac{1}{4^n}$. Intuitively, the curve spends the same "amount of time" travelling through each square of the grid. Now we proceed in proving the lemma.

We observe that when we plot H_n and divide the grid into 4^n sub-squares with side length $\frac{1}{2^n}$, each square contains exactly one copy of the original Π shaped motif from H_0. Then plotting H_{n+1} and increasing the resolution of the grid to contain 4^{n+1} squares of side length $\frac{1}{2^{n+1}}$, we observe each of the original larger sub-squares are replaced by their four quadrants which are visited by H_{n+1} one after another. It is easy to see that each collection of four sub-squares will also be visited in the same order as the larger squares they make up. From this we can conclude that for any interval $I \subseteq [0,1]$ such that $H_n(I)$ is a subset of one of the larger sub-squares of side length $\frac{1}{2^n}$ then $H_{n+1}(I)$ must be a subset of the union of the four squares of side length $\frac{1}{2^{n+1}}$ which make up the original larger square. Thus for any $x \in [0,1]$, $H_n(x)$ and $H_{n+1}(x)$ must fall in the same square with side length $\frac{1}{2^n}$, thus the furthest apart they could be is $\frac{\sqrt{2}}{2^n}$. Thus $||H_n - H_{n+1}||_\infty \leq \frac{\sqrt{2}}{2^n}$. □

Theorem 3. *The sequence $(H_n)_{n\in\mathbb{N}}$ is Cauchy so converges uniformly.*

Proof. The metric space $C(([0,1],[0,1]^2), ||\cdot||_\infty)$ of continuous functions from $[0,1]$ to $[0,1]^2$ with the infinity norm is complete [4] (pg. 37-38). We proceed by showing the sequence $(H_n)_{n\in\mathbb{R}}$ is Cauchy in $C(([0,1],[0,1]^2), ||\cdot||_\infty)$.

Let $\varepsilon > 0$, we can choose $N \in \mathbb{N}$ such that $\frac{\sqrt{2}}{2^{N-1}} < \varepsilon$ and for $m > n > N$ we then have:

$$\begin{aligned}||H_n - H_m||_\infty &\leq ||H_n - H_{n+1}||_\infty + ||H_{n+1} - H_{n+2}||_\infty + ||H_{n+2} - H_{n+3}||_\infty + ... + ||H_{m-1} - H_m||_\infty \\ &\leq \frac{\sqrt{2}}{2^n} + \frac{\sqrt{2}}{2^{n+1}} + \frac{\sqrt{2}}{2^{n+2}} + ... + \frac{\sqrt{2}}{2^m} \leq \frac{\sqrt{2}}{2^n} + \frac{\sqrt{2}}{2^{n+1}} + \frac{\sqrt{2}}{2^{n+2}} + ... \\ &= \sqrt{2}\left(\frac{1}{2^n} + \frac{1}{2^{n+1}} + \frac{1}{2^{n+2}} + ...\right) = \frac{\sqrt{2}}{2^{n-1}} < \varepsilon\end{aligned}$$

Since $C(([0,1],[0,1]^2),||\cdot||_\infty)$ is a complete metric space, the sequence $(H_n)_{n\in\mathbb{N}}$ converges uniformly to a continuous function. □

It is the continuous function which the sequence $(H_n)_{n\in\mathbb{N}}$ converges to uniformly which we define to be the Hilbert curve, H.

3.3 The Hilbert curve is space-filling

From the definition given above, it is not intuitively obvious that the Hilbert curve is in fact a space-filling curve. It is clear to see that the successive approximation curves seem to get progressively more "dense" (see Figure 5), however each approximation curve H_n still only travels through an infinitely small sliver off all the points in $[0,1]^2$. In this section we verify that the Hilbert curve which we rigorously defined in the previous section does in fact meet the requirements to be considered a space-filling curve; namely, we verify that it is surjective.

Firsly we prove a lemma concerning convergent sequences within closed sets.

Lemma 11. *For a closed set $A \in \mathbb{R}^n$, if $(x_n)_{n\in\mathbb{N}}$ is a convergent sequence where for all n, $x_n \in A$, then $(x_n)_{n\in\mathbb{N}}$ converges in A.*

Proof. Let us assume $\lim\limits_{n\to\infty}(x_n) = x$ is not an element of A, then $x \in A^c$ which is an open set. Therefore there exists $\varepsilon > 0$ such that $B(x,\varepsilon) \subseteq A^c$. By the definition of convergence, there exists $N \in \mathbb{N}$ such that for all $n > N$, $|x_n - x| < \varepsilon$, however this implies that $x_n \in B(x,\varepsilon) \subseteq A^c$ which contradicts the fact that $x_n \in A$ for all n. Hence the original assumption that the limit of the sequence is not an element of A must be false. □

Next, we introduce the notion of sequential compactness, which will be helpful later in our proof.

Definition 6. *A set A is sequentially compact if every sequence $(x_n)_{n\in\mathbb{N}}$ in A converges in A.*

Lemma 12. $[0,1]$ *is sequentially compact.*

Proof. It is easy to see $[0,1]$ is closed and bounded. By the Bolzano–Weierstrass theorem, every bounded sequence of real numbers has a convergent subsequence [3] (pg. 78-79). Any sequence $S = (x_n)_{n\in\mathbb{N}}$ in $[0,1]$ must therefore have a convergent subsequence $S' = (x_{n_t})_{n\in\mathbb{N}}$. Since each term of S' is in $[0,1]$ and $[0,1]$ is closed, S' must converge in $[0,1]$, therefore $[0,1]$ is sequentially compact. □

The final preliminary result we require before proving the Hilbert curve's surjectivity is a version of Cantor's intersection theorem.

Theorem 4 (Cantor's Intersection Theorem)**.** *Given a sequence of non-empty, closed, nested subsets of $\mathbb{R}^n$, $(C_n)_{n\in\mathbb{N}}$ whose diameters tend to zero, $\lim\limits_{n\to\infty}\mathrm{Diam}(C_n) = 0$, then their intersection contains exactly one point:*

$$\bigcap_{n=1}^{\infty} C_n = \{x\}$$

Proof. Since the diameters of the sets in the sequence tends to zero, the intersection of all the sets must have diameter zero. A set with diameter zero must either contain a single point or no points at all; we show that the intersection cannot be empty. By choosing an element x_n from each set in the sequence C_n, we get a sequence $(x_n)_{n\in\mathbb{N}}$. Given $\varepsilon > 0$, there must be $N \in \mathbb{N}$ such that for all $n > N$, $\mathrm{Diam}(C_n) < \varepsilon$, and since the sets are nested $|x_n - x_m| < \varepsilon$ for all $n, m > N$, so the sequence $(x_n)_{n\in\mathbb{N}}$ is Cauchy. Since $\mathbb{R}^n$ with the Euclidean distance metric is a complete metric space, this sequence must converge to some point x. Since each of the sets is closed, and x is the limit of a sequence in C_n, x must be in C_n for all $n \in \mathbb{N}$. Finally, since this is true for all C_n, x must be included in the intersection of all the sets, therefore the intersection is not empty, instead it contains the single point x.

$$\bigcap_{n=1}^{\infty} C_n = \{x\}$$

□

With these preliminaries covered, we can now proceed in proving that the Hilbert curve is surjective.

Theorem 5. *The Hilbert curve is surjective:* $H([0,1]) = [0,1]^2$.

Proof. Let $(x,y) \in [0,1]^2$. We wish to find $t \in [0,1]$ such that $H(t) = (x,y)$. We note it is possible to construct a sequence of progressively smaller square subsets $(C_n)_{n\in\mathbb{N}}$ of the unit square which always contains (x,y), where the nth square is taken from the unit square partitioned into 4^n sub-squares.

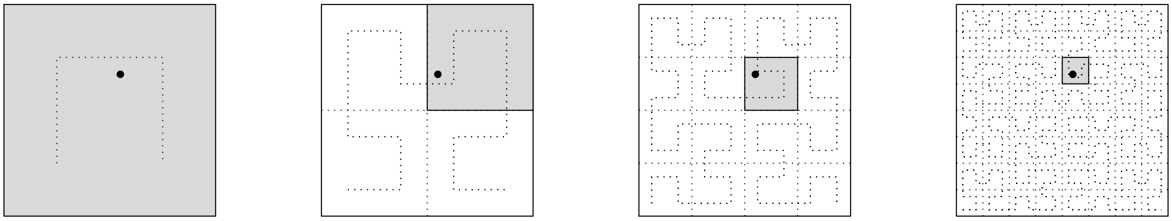


Figure 6: The first four elements in the sequence $(C_n)_{n\in\mathbb{N}}$; C_0, C_1, C_2, C_3, all containing the same arbitrary point. H_n is also showed overlayed on the square containing C_n.

We observe that since these sets are closed, nested and have diameter tending to zero in the complete metric space of $\mathbb{R}^2$ with the Euclidean distance metric and that each of them contains (x,y), that by Cantor's intersection theorem $\bigcap\limits_{n=0}^{\infty} C_n = \{(x,y)\}$.

It can also be seen that by virtue of the Hilbert curve's construction, there will always be a section of H_n which passes through C_n. Thus there is a sequence $(t_n)_{n\in\mathbb{R}}$ such that $H_n(t_n) \in C_n$. Since $[0,1]$ is sequentially compact, there must exist a subsequence $(t_{n_k})_{n\in\mathbb{R}}$ which converges to some $t \in [0,1]$. We can see that $(x,y) = \lim\limits_{n\to\infty} H_{n_k}(t_{n_k})$, what remains is to establish that $\lim\limits_{n\to\infty} H_{n_k}(t_{n_k}) = H(t)$

Let $\varepsilon > 0$, since H is continuous, there exists $\delta > 0$ such that $|t-t'| < \delta$ implies $|H(t) - H(t')| < \frac{\varepsilon}{2}$. Since $t_{n_k} \to t$, there exists $P \in \mathbb{N}$ such that for all $n > P$, $|t - t_{n_k}| < \delta$ implies $|H(t) - H(t_{n_k})| < \frac{\varepsilon}{2}$. Since $H_n \to H$ uniformly, there exists $Q \in \mathbb{N}$ such that for all $n > Q$, $||H_n - H||_\infty < \frac{\varepsilon}{2}$. Let $N = \max(P,Q)$, then for all $n > N$:

$$\begin{aligned}
|H_n(t_{n_k}) - H(t)| &= |H(t_{n_k}) + (H(t_{n_k}) - H(t_{n_k})) - H(t)| \\
&= |(H_n(t_{n_k}) - H(t_{n_k})) + (H(t_{n_k}) - H(t))| \\
&\leq |H_n(t_{n_k}) - H(t_{n_k})| + |H(t) - H(t_{n_k})| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon
\end{aligned}$$

So $\lim\limits_{n\to\infty} H_n(t_{n_k}) = H(t) = (x,y)$, and so H is surjective. □

Since H is surjective, we can conlude that the Hilbert curve is in fact a space-filling curve.

3.4 The Hölder Continuity of the Hilbert curve

Space-filling curves have several contemporary applications. For example, during the implementation of a piece of software like Google Maps, a natural question which arises is "How can we tile the earth and index every tile such that neighboring tiles can be loaded from the computer's memory in an efficient manner?". Such a task requires mapping from one dimension (computer memory) to two dimensions (a map of the world) in such a way that points close by in one dimension are mapped close to each other in two dimensions too. In fact it is specifically the Hilbert curve, or more precisely, due to the obvious physical limitations of computers, an approximation of the Hilbert curve (which also has the advantage of being bijective), which is used by Google to make this mapping [1].

Another application of space-filling curves is in designing efficient algorithms for executing matrix operations. Due to the construction of modern computers, the speed which data is retrieved from memory by the computer's central processing unit (CPU) does not always take the same amount of time. Memory caches store data which has been recently used by the CPU in such a way that it can be very quickly accessed again. This means algorithms can be made more efficient by strategically batching operations which use the same or similar data. By using the same property of the Hilbert curve levaraged by Google Maps, operations like matrix-vector multiplication can be optimised considerably [2] (pg. 195-214).

Both of these applications make use of the locality preserving properties of the Hilbert curve, that is, the ability of the Hilbert curve to provide a relatively good distance preserving map between one and two dimensions. Given two numbers $x_1, x_2 \in [0,1]$ "close together", we can be sure their images $H(x_1)$ and $H(x_2)$ will also be relatively "close together". The notion of Hölder continuity can be used to more rigorously quantify this property.

Definition 7 (Hölder continuous)**.** *A function $f : I \to \mathbb{R}^n$ on the interval $I \subseteq \mathbb{R}^m$ is said to be Hölder continuous with exponent r over I, if there is a constant $C > 0$ such that for all $x, y \in I$*

$$|f(x) - f(y)| \leq C|x - y|^r$$

Theorem 6. *The Hilbert curve is Hölder continuous for prcisely the exponents $0 \leq r \leq \frac{1}{2}$, where $r \in \mathbb{R}$.*

Proof. If we split the unit square $[0,1]^2$ into 4^n squares of side length 2^{-n}, we observe that by construction H_n passes through each of these squares, mapping an interval $I \subset [0,1]$ of diameter exactly 4^{-n} onto each of these squares. We also note adjacent intervals are mapped to adjacent squares. Let $x, y \in [0,1]$, we choose $n \in \mathbb{N}$ such that $4^{-(n+1)} \leq |x-y| < 4^{-n}$. Since $|x - y| < 4^{-n}$, x and y must be in the same or adjacent intervals of $[0,1]$ and so $H_n(x)$ and $H_n(y)$ must be in the same or adjacent squares of sde length 2^{-n}, let S_x be the square containing x and S_y that containing y. By construction of the Hilbert approximation curves we observe that if $H_n(x) \in S_x$ then also $H_{n+1}(x) \in S_x$, $H_{n+2}(x) \in S_x$, etc., since $(H_n)_{n \in \mathbb{N}}$ converges pointwise (since uniform convergence entails pointwise convergence) and S_x is closed, $\lim_{n\to\infty} H_n(x) = H(x) \in S_x$ and similarly $H(y) \in S_y$. The maximum distance between $H(x)$ and $H(y)$ is therefore bounded by the length of the diagonal of the rectangle formed by two adjacent squares, that length being:

$$2^{-n}\sqrt{1^2 + 2^2} = 2^{-n}\sqrt{5}$$

Hence we get:

$$\begin{aligned} |H(x) - H(y)| &\leq 2^{-n}\sqrt{5} = (4^{-n})^{\frac{1}{2}}\sqrt{5} = (4^{-(n+1)})^{\frac{1}{2}}4^{\frac{1}{2}}\sqrt{5} \\ &\leq 2\sqrt{5}|x-y|^{\frac{1}{2}} \end{aligned}$$

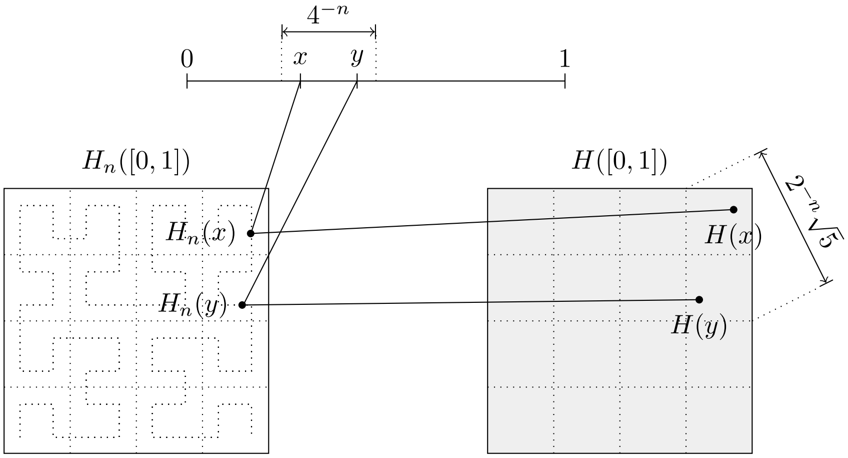


Figure 7: Geometric visualisation of the derived bounds proving Hölder continuity of the Hilbert curve

Thus the Hilbert curve is Hölder continuous with exponent $\frac{1}{2}$. Note that since $x, y \in [0,1]$, $|x - y| \leq 1$ and so for all r, $0 \leq r < \frac{1}{2}$, $|x-y|^r > |x-y|^{\frac{1}{2}}$, therefore the curve is also Hölder continuous for all other exponents between zero and $\frac{1}{2}$. What remains is to show that $\frac{1}{2}$ is the greatest exponent for which the Hilbert curve is Hölder continuous.

Assume there exists $r > \frac{1}{2}$ for which the Hilbert curve is Hölder continuous, then it follows that:

$$\frac{|H(x) - H(y)|}{|x - y|^r} \leq C$$

for $C \in \mathbb{R}_{\geq 0}$. We proceed by showing there is a sequence of $((x_n, y_n))_{n \in \mathbb{N}}$ such that

$$\lim_{n \to \infty} \frac{|H(x_n) - H(y_n)|}{|x_n - y_n|^r} = \infty$$

Contradicting the assertion that the left-hand side of the inequality is bounded.

We choose $x_n, y_n \in [0, 1]$ such that $H_n(x_n)$ and $H_n(y_n)$ are mapped to the start and end of one of the $\frac{1}{2^n}$ line segments used to construct H_n *which is not* in the bottom left or bottom right sub-square of the construction (as illustrated in Figure 8). The line segment of length $\frac{1}{2^n}$ will make up $\frac{1}{4^n - 1}$ of the total length of the Hilbert curve, so $|x_n - y_n| \leq \frac{1}{4^n - 1}$, note that if the Hilbert curve was traversed at a constant rate, this would be a strict equality, but since the part of the curve in the bottom left and bottom right sub-squares is traversed at a "slower" rate (in order to preserve a constant sub-square traversal rate), it must move "faster" through the others. From the above, we have that

$$\begin{aligned}\lim_{n \to \infty} \frac{|H(x_n) - H(y_n)|}{|x_n - y_n|^r} &\geq \lim_{n \to \infty} \frac{2^{-n}}{(4^n - 1)^{-r}} = \lim_{n \to \infty} \frac{(4^n - 1)^r}{2^n} \\ &= \lim_{n \to \infty} \frac{4^{nr}}{2^n} = \lim_{n \to \infty} \frac{2^{2nr}}{2^n} = \lim_{n \to \infty} 2^{n(2r-1)}\end{aligned}$$

Since $r > \frac{1}{2}$, the $2r - 1$ will be positive and so $\lim\limits_{n \to \infty} 2^{n(2r-1)}$ will tend to infinity, contradicting the assertion that $\frac{|H(x_n) - H(y_n)|}{|x_n - y_n|^r}$ is bounded. □

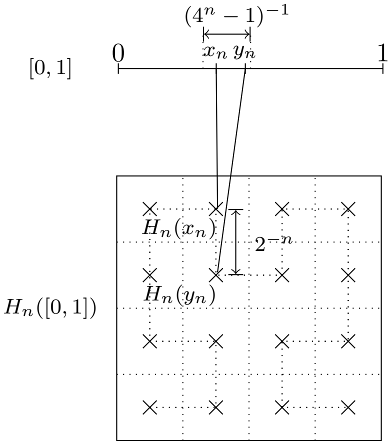


Figure 8: Geometric visualisation of relationship between elements of the sequence $((x_n, y_n))_{n \in \mathbb{N}}$ and elements of the sequence $((H_n(x_n), H_n(y_n))_{n \in \mathbb{N}}$

4 Self-similarity and fractal dimension of Hilbert curve coordinate functions

4.1 Introduction to self-similarity and iterated function systems

Self-similarity is the property possessed by an object which resembles itself at a variety of different scales. Coastlines, the silhouettes of mountain ranges, and fern leaves are all real world examples of entities which exhibit self-similar behaviour (obviously, due to the limits imposed by physical laws, such similarities do not continue to indefinitely small scales). Some mathematical shapes exhibiting self similar properties include the Sierpiński triangle, the Mandelbrot set, and as we will explore, the coordinate functions of the Hilbert curve. We can more rigorously and formally understand what it means for a set to be self similar using mathematical language.

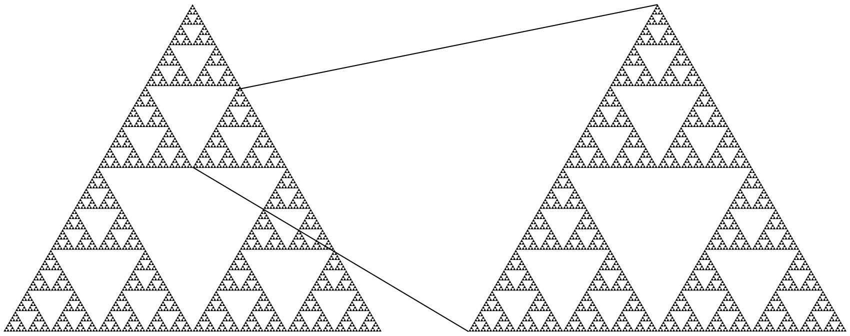


Figure 9: The Sierpiński triangle is an example of a self-similar mathematical object: Each triangular sub-section of the shape is similar to the shape as a whole.

Definition 8. *A function $f : \mathbb{R}^n \to \mathbb{R}^n$ is a similarity if there is a positive real number r such that for all $x, y \in \mathbb{R}^n$, $|f(x) - f(y)| = r|x - y|$. If $r < 1$, the similarity is a contractive similarity.*

Definition 9 (Iterated function system)**.** *A finite collection of contractive similarities $\{f_i\}_{i=1}^m$ is called an iterated function system (IFS).*

Given an iterated function system, there is a unique, non-empty compact set $A \subset \mathbb{R}^n$ such that:

$$A = \bigcup_{i=1}^{m} f_i(A)$$

This result is proved as **Theorem 4.1.3** in [6]. The set A is said to be the *invariant* set of the IFS and it is these invariant sets which are defined to be self-similar. The notion of an IFS can be generalised to that of a *graph directed IFS*

Definition 10. *A graph directed IFS in $\mathbb{R}^n$ consists of a directed multigraph (V, E), with a finite set of vertices V joined by a finite set of directed edges E where each edge $e \in E$ is associated with a contractive similarity $f_e : \mathbb{R}^n \to \mathbb{R}^n$. We denote the set of all edges from the vertex v to the vertex u as E_{vu}.*

Generalising the notion of the invariant of an IFS, for a graph directed IFS there exists a unique list of non-empty compact sets $(A_v | v \in V)$ such that for every $u \in V$:

$$A_u = \bigcup_{v \in V, e \in E_{uv}} f_e(A_v) \tag{1}$$

This result is proved as **Theorem 4.3.5.** in [6]. Note a "regular" IFS can be seen as just a digraph IFS where $|V| = 1$. Using the notion of graph directed IFS we have a framework we can use to investigate the coordinate functions of the Hilbert curve.

4.2 Graph directed IFS and the Hilbert curve coordinate functions

The self-referential recursive nature of the Hilbert curve's construction results in the curve's coordinate functions also exhibiting interesting self-similar, fractal properties. In this and the following section, we will investigate these properties.

Definition 11 (Coordinate functions)**.** *The Hilbert curve coordinate functions are the pair of functions $f(t)$ and $g(t)$ such that $H(t) = (f(t), g(t))$.*

In this section, we will construct a graph directed IFS and then show its invariant sets are equal to the graphs of the Hilbert curve coordinate functions f and g. Let A and B be the matrices:

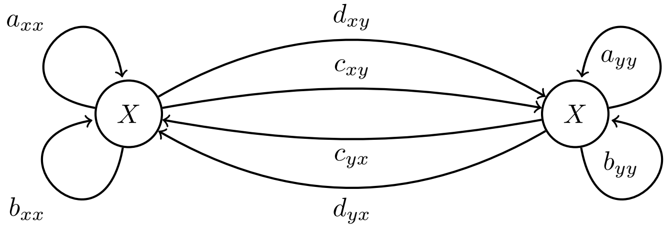
$$A = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \qquad B = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}$$

We then define the following affine functions, which will each be associated with an edge of the graph directed IFS.

$$a_{xx}(\mathbf{x}) = A\mathbf{x} + \begin{pmatrix} \frac{1}{4} \\ 0 \end{pmatrix} \qquad a_{yy}(\mathbf{x}) = A\mathbf{x} + \begin{pmatrix} \frac{1}{4} \\ 0 \end{pmatrix}$$

$$\begin{aligned} b_{xx}(\mathbf{x}) &= A\mathbf{x} + \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} & b_{yy}(\mathbf{x}) &= A\mathbf{x} + \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} \\ c_{xy}(\mathbf{x}) &= A\mathbf{x} & c_{yx}(\mathbf{x}) &= A\mathbf{x} \\ d_{xy}(\mathbf{x}) &= A\mathbf{x} + \begin{pmatrix} \frac{3}{4} \\ 1 \end{pmatrix} & d_{yx}(\mathbf{x}) &= A\mathbf{x} + \begin{pmatrix} \frac{3}{4} \\ \frac{1}{2} \end{pmatrix} \end{aligned}$$

Below is an illustration of the graph directed IFS. We also label the two vertices in our directed graph as X and Y to more easily disambiguate between them.



Theorem 7. *The invariant sets A_X and A_Y of the vertices X and Y of the graph directed IFS defined above equal the graphs of the coordinate functions of the Hilbert curves, defined as* $\textsc{Graph}(f) = \{(t, f(t)) | t \in [0,1]\}$, $\textsc{Graph}(g) = \{(t, g(t)) | t \in [0,1]\}$. *Specifically* $A_X = \textsc{Graph}(f)$ *and* $A_Y = \textsc{Graph}(g)$.

Proof. To prove this result, we show that the image of each of the given affine functions of $\textsc{Graph}(f)$ or $\textsc{Graph}(g)$ is a subset of either $\textsc{Graph}(f)$ or $\textsc{Graph}(g)$. For example, for the affine function $a_{xx}(\mathbf{x})$, we wish to show that $a_{xx}(\textsc{Graph}(f)) \subseteq \textsc{Graph}(f)$.

Let $(t, f(t))$ where $t \in [0,1]$ be an arbitrary point in $\textsc{Graph}(f)$, we wish to show $a_{xx}((t, f(t))) \in \textsc{Graph}(f)$, i.e. $a_{xx}((t, f(t))) = (t', f(t'))$ for some $t' \in [0,1]$.

$$a_{xx}((t, f(t)) = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} t \\ f(t) \end{pmatrix} + \begin{pmatrix} \frac{1}{4} \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{4}t + \frac{1}{4} \\ \frac{1}{2}f(t) \end{pmatrix}$$

By considering the geometry of the Hilbert curve, we can deduce that $\frac{1}{2}f(t) = f(\frac{1}{4}t + \frac{1}{4})$: We observe that since $\frac{1}{4}t + \frac{1}{4} \in [\frac{1}{4}, \frac{1}{2}]$, then $H(\frac{1}{4}t + \frac{1}{4})$ will be in the top left quadrant of the unit square ($H([\frac{1}{4}, \frac{1}{2}])$.

Since this top left quadrant is similar to the whole of the image of H (it is scaled by a factor of a half), the point $H(\frac{1}{4}t+\frac{1}{4})$ can be found at the same point relative to the top left quadrant of the unit square as $H(t)$ can be found with respect to the whole unit square. Aided with a visualisation (below) it is straightforward to see that this implies $f(\frac{1}{4}t+\frac{1}{4})=\frac{1}{2}f(t)$.

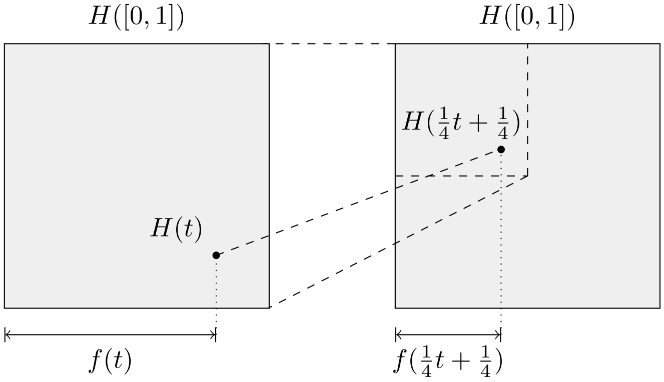


Figure 10: Geometric visualisation showing why $f(\frac{1}{4}t+\frac{1}{4})$ is half of $f(t)$

Hence

$$\begin{pmatrix} \frac{1}{4}t+\frac{1}{4} \\ \frac{1}{2}f(t) \end{pmatrix} = \begin{pmatrix} \frac{1}{4}t+\frac{1}{4} \\ f(\frac{1}{4}t+\frac{1}{4}) \end{pmatrix} = \begin{pmatrix} t' \\ f(t') \end{pmatrix} \in \textsc{Graph}(f)$$

So we can see that $a_{xx}(\textsc{Graph}(f)) \subseteq \textsc{Graph}(f)$. More precisely, since for all $t \in [0,1]$, $a_{xx}((t, f(t)) = (t', f(t'))$ we know that $t' \in [\frac{1}{4}, \frac{1}{2}]$ then $a_{xx}(\textsc{Graph}(f)) = \{(t, f(t)|t \in [\frac{1}{4}, \frac{1}{2}]\} \subset \textsc{Graph}(f)$.

As for a_{xx}, for each of the affine functions given, there is an associated equation in terms of the coordinate functions which we can show holds. We can use a similar approach as the one given above for the case of a_{xx} to derive each of these equations and determine the set which is the image of each of these affine functions.

Affine function	Equation	Image
a_{xx}	$f(\frac{1}{4}t+\frac{1}{4})=\frac{1}{2}f(t)$	$a_{xx}(\textsc{Graph}(f))=\{(t,f(t))\|t\in[\frac{1}{4},\frac{1}{2}]\}$
b_{xx}	$f(\frac{1}{4}t+\frac{1}{2})=\frac{1}{2}+\frac{1}{2}f(t)$	$b_{xx}(\textsc{Graph}(f))=\{(t,f(t))\|t\in[\frac{1}{2},\frac{3}{4}]\}$
c_{xy}	$f(\frac{1}{4}t)=\frac{1}{2}g(t)$	$c_{xy}(\textsc{Graph}(g))=\{(t,f(t))\|t\in[0,\frac{1}{4}]\}$
d_{xy}	$f(\frac{1}{4}t+\frac{3}{4})=1-\frac{1}{2}g(t)$	$d_{xy}(\textsc{Graph}(g))=\{(t,f(t))\|t\in[\frac{3}{4},1]\}$
a_{yy}	$g(\frac{1}{4}t+\frac{1}{4})=\frac{1}{2}+\frac{1}{2}g(t)$	$a_{yy}(\textsc{Graph}(g))=\{(t,g(t))\|t\in[\frac{1}{4},\frac{1}{2}]\}$
b_{yy}	$g(\frac{1}{4}t+\frac{1}{2})=\frac{1}{2}+\frac{1}{2}g(t)$	$b_{yy}(\textsc{Graph}(g))=\{(t,g(t))\|t\in[\frac{1}{2},\frac{3}{4}]\}$
c_{yx}	$f(\frac{1}{4}t)=\frac{1}{2}g(t)$	$c_{yx}(\textsc{Graph}(f))=\{(t,g(t)\|t\in[0,\frac{1}{4}]\}$
d_{yx}	$f(\frac{1}{4}t+\frac{3}{4})=\frac{1}{2}-\frac{1}{2}g(t)$	$d_{yx}(\textsc{Graph}(f))=\{(t,g(t)\|t\in[\frac{3}{4},1]\}$

And so we see that:

$$\textsc{Graph}(f) = a_{xx}(\textsc{Graph}(f)) \cup b_{xx}(\textsc{Graph}(f)) \cup c_{xy}(\textsc{Graph}(g)) \cup d_{xy}(\textsc{Graph}(g))$$

$$\textsc{Graph}(g) = a_{yy}(\textsc{Graph}(g)) \cup b_{yy}(\textsc{Graph}(g)) \cup c_{yx}(\textsc{Graph}(f)) \cup d_{yx}(\textsc{Graph}(f))$$

Thus $\textsc{Graph}(f)$ and $\textsc{Graph}(g)$ satisfy the properties required of the unique invariant sets A_X and A_Y (equation (1)) for the graph directed IFS. Since the invariant sets are unique, it must be the case that $X = \textsc{Graph}(f)$ and $Y = \textsc{Graph}(g)$. □

Below are the graphs of the two invariant sets of the graph directed IFS, which we have shown correspond to the graphs of the coordinate functions of the Hilbert curve f and g.

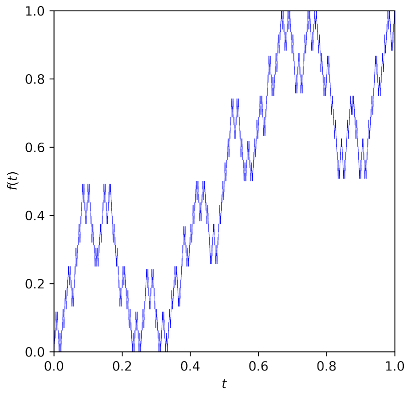


Figure 11: Graph of f

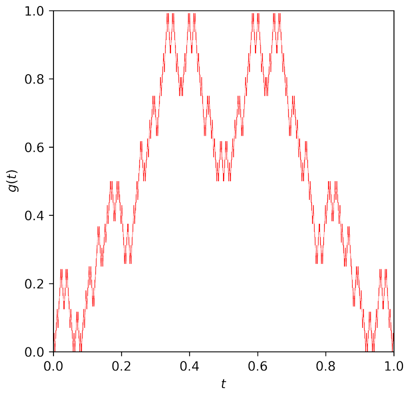


Figure 12: Graph of g

4.3 The box counting dimension of the Hilbert curve coordinate functions

The fractal dimension of a shape is a positive real number (not necessarily an integer) which describes how the complexity of a shape changes with respect to the scale at which it is measured at. For "ordinary" geometric shapes like lines, squares, triangles, spheres, etc. the shape's fractal dimension will coincide with its spatial dimension [14], however as we will see, for self similar shapes like the Hilbert curve coordinate functions, the fractal dimension may not be an exact integer. There are many different types of fractal dimension, with different corresponding definitions. One of the more straightforward types of fractal dimension and the one we will discuss here is the *box counting dimension*. In this section we will define the box dimension and determine the box counting dimension of both of the coordinate functions of the Hilbert curve.

Definition 12 (Box counting dimension)**.** *Given a bounded set $F \in \mathbb{R}^n$, a finite or countable collection of sets $\{U_i\}_i$ is said to be a δ-cover for F if $F \subseteq \bigcup_i U_i$ and $\operatorname{diam}(U_i) \leq \delta$ for all i. For a set F, let $N_\delta(F)$ be the smallest possible number of sets in a δ-cover for F.*

$$\textit{The lower box dimension of } F \textit{ is given by } \underline{\dim}_B F = \liminf_{\delta \to 0} \frac{\log N_\delta(F)}{-\log(\delta)}$$

$$\textit{The upper box dimension of } F \textit{ is given by } \overline{\dim}_B F = \limsup_{\delta \to 0} \frac{\log N_\delta(F)}{-\log(\delta)}$$

If $\overline{\dim}_B F = \underline{\dim}_B F$ then their common value is taken to be the box counting dimension of F, $\dim_B F$. If this limit exists, there are several different definitions of $N_\delta(F)$ which can be taken and yield the same value for $\dim_B F$. One such alternative definition which we will make use of is to take $N_\delta(F)$ to be the number of squares in the $\delta-$grid, which is the grid formed by the set of lines parallel to the coordinate axes spaced apart by a distance δ, which the set F intersects, this equivalence is proved in [7](pg. 43-45).

Having defined the box counting dimension, we proceed in finding the value of the box counting dimension of the Hilbert curve coordinate functions.

Theorem 8. *The graphs of both of the Hilbert curve coordinate functions* Graph(f) *and* Graph(g) *have box dimension equal to $\frac{3}{2}$.*

Proof. Instead of considering the limits of the relevant expressions as $\delta \to 0$ continuously, we can consider the limit along a sequence of discrete values $(c^n)_{n \in \mathbb{N}}$ for some $c \in (0, 1)$, this is shown in [7] (pg. 44-45). We choose to take the limit along the sequence $\{1, \frac{1}{4^1}, \frac{1}{4^2}, ...\}$.

To determine how many boxes of the δ-grid the curve fills, we begin by considering two copies of the unit square, with one being associated with each vertex in the directed graph, and we observe how the graph directed IFS acts on these squares. Successive iterations of the IFS on these unit squares can be seen in Figure 13. Since both the invariant sets Graph(f) and Graph(g) are subsets of the unit square, by induction, each time we apply the IFS, the resultant sets must still contain these invariant sets. After each application of the IFS, we generate two set of progressively smaller and more numerous rectangles, and note $\textsc{Graph}(f)$ is covered by and passes through each rectangle in one of these sets, and the same is the case for $\textsc{Graph}(g)$ and the other of these sets. More precisely, after the IFS has acted n times, we have associated with each vertex u a set $B_u^{(n)}$

$$B_u^{(n)} = \bigcup_{\substack{|\mathbf{e}|=n \\ i(\mathbf{e})=u}} = f_{\mathbf{e}}(B_{t(\mathbf{e})}^{(0)})$$

Where for the graph's set of vertices V and edges E, the functions $i : E \to V$, $t : E \to V$ are the *initial* and *terminal* vertex functions, $B_v^{(0)}$ is the initial set associated with the vertex v which the IFS acts on, $\mathbf{e}$ is a list of edges $e_1 e_2 ... e_n$ where $t(e_i) = i(e_{i+1})$ and finally $f_{\mathbf{e}} = f_{e_1} \circ f_{e_2} \circ ... \circ f_{e_n}$. We note that

$$\textsc{Graph}(f) \subseteq B_X^{(n)} = \bigcup_{\substack{|\mathbf{e}|=n \\ i(\mathbf{e})=X}} = f_{\mathbf{e}}([0,1]^2)$$

and

$$\textsc{Graph}(g) \subseteq B_Y^{(n)} = \bigcup_{\substack{|\mathbf{e}|=n \\ i(\mathbf{e})=Y}} = f_{\mathbf{e}}([0,1]^2)$$

for all $n \in \mathbb{N}$. We can use these sets of rectangles $B_u^{(n)}$ which cover the invariants sets to determine $N_{4^{-n}}(\textsc{Graph}(f))$ and $N_{4^{-n}}(\textsc{Graph}(g))$.

It is easy to see that for each iteration the number of rectangles in $B_X^{(n)}$ and $B_Y^{(n)}$ increases by a factor of four, and at the nth iteration, there are 4^n rectangles and each rectangle has width 4^{-n} and height 2^{-n} so consist of 2^n squares of width 4^{-n}, and so $N_{4^{-n}}(\textsc{Graph}(f)) = N_{4^{-n}}(\textsc{Graph}(g)) = 4^n \cdot 2^n = 8^n$. Since for $n \geq 1$

$$\frac{\log N_{4^{-n}}(\textsc{Graph}(g)))}{-\log(4^{-n})} = \frac{\log N_{4^{-n}}(\textsc{Graph}(g)))}{\log(\frac{1}{4^{-n}})} = \frac{\log(8^n)}{\log(4^n)} = \frac{\log(2^{3n})}{\log(2^{2n})} = \frac{3}{2}$$

is a constant, then the upper and lower limits trivially exist and coincide for the sequences

$$\Big(\frac{\log N_{4^{-n}}(\textsc{Graph}(f)))}{-\log(4^{-n})}\Big)_{n\in\mathbb{N}} \quad \text{and} \quad \Big(\frac{\log N_{4^{-n}}(\textsc{Graph}(g)))}{-\log(4^{-n})}\Big)_{n\in\mathbb{N}}$$

and so finally we can conclude that:

$$\dim_B(\textsc{Graph}(f)) = \dim_B(\textsc{Graph}(g)) = \lim_{n\to\infty} \frac{\log(8^n)}{-\log(4^{-n})} = \lim_{n\to\infty} \frac{\log(2^{3n})}{\log(2^{2n})} = \frac{3}{2}$$

□

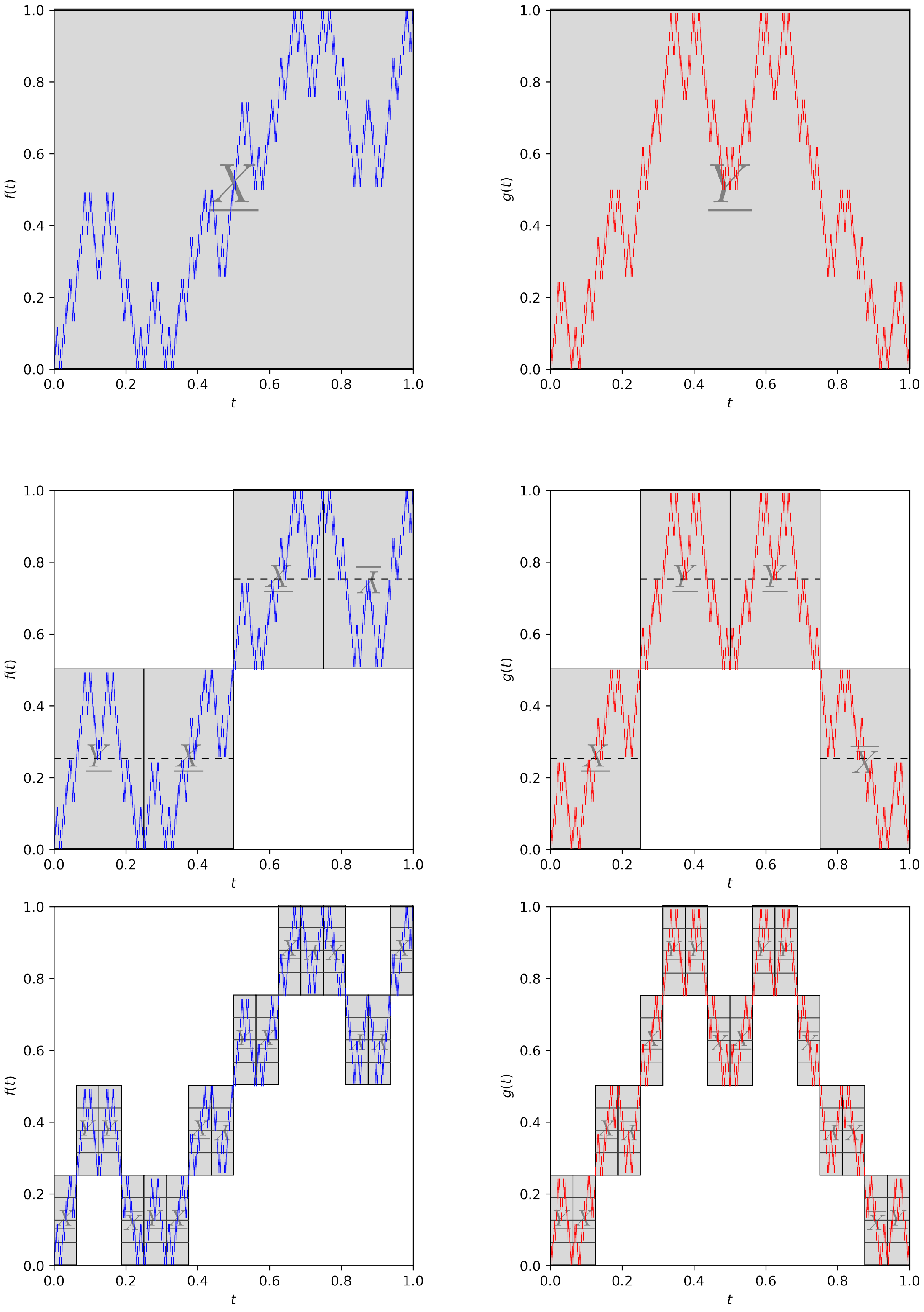


Figure 13: Successive iterations of the IFS on the two unit squares yield sets $B_X^{(n)}$ and $B_Y^{(n)}$ which cover the graphs of the Hilbert coordinate functions and can be subdivided into squares of a $\frac{1}{2^n}$-grid. In the top two images, the unit square associated with the vertex X is labelled $\underline{X}$ and the unit square for the vertex Y is similarly labelled $\underline{Y}$, the effect of each of the affine transformations on these two unit squares can be seen in successive iterations of the IFS in the lower images.

5 Outlook

This report has given an overview of the topic of space-filling curves, including a discussion of historical background, an investigation of one of the most important theorems relating to space-filling curves, and an in depth discussion and analysis of a specific space-filling curve. However this report does not represent and exhaustive survey of space-filling curves in general, or even for the specific example of the Hilbert curve; there are several directions further work building on that done in this project could be taken, and topics which the interested reader might appreciate being directed towards. For example, in the discussion of the Hölder continuity of the Hilbert curve, we derive the following relationship:

$$\frac{|H(x)-H(y)|}{|x-y|^{\frac{1}{2}}} \leq C$$

Which taking the logarithm of both sides and rearranging, gives

$$\log|H(x)-H(y)| \leq \log C + \frac{1}{2}\log|x-y|$$

implying

$$\frac{\log|H(x)-H(y)|}{\log|x-y|} \leq \frac{\log C}{\log|x-y|} + \frac{1}{2}$$

so

$$\limsup_{y\to x} \frac{\log|H(x)-H(y)|}{\log|x-y|} \leq \frac{1}{2}$$

The right-hand side of this expression defines the upper local dimension of H at x, $\overline{\dim}_{\text{loc}}H(x)$ (with the lower local dimension being defined similarly, except taking the lower limit rather than the upper) [8] (pg. 169). The local dimension of functions is a quantity studied in the field of multi-fractal geometry, and it is generally believed that naturally occurring functions (such as functions whose graphs are the invariant sets of graph directed IFS's) have constant local dimension d almost everywhere [12][13] (that is to say for an appropriate measure μ, the set $N = \{x|\dim_{\text{loc}} f(x) \neq 0\}$ has measure zero, $\mu(N) = 0$). From this, combined with what we already know about the Hilbert curve from work done in this project, it is reasonable to conjecture that

$$\overline{\dim}_{\text{loc}}H(x) = \underline{\dim}_{\text{loc}}H(x) = \dim_{\text{loc}} H(x) = \frac{1}{2}$$

for x almost everywhere in [0, 1].

In this report, the box counting dimension of the coordinate functions of the Hilbert curve were discussed, however as mentioned, the box dimension is only one of many definitions of fractal dimension, each of which has different properties and use cases. For example, another commonly used fractal dimension is Hausdorff dimension. We define

$$H^d_\delta(S) = \inf\left\{\sum_{i=1}^{\infty}(\operatorname{diam} U_i)^d | \{U_i\}_{i\in\mathbb{N}} \text{ is a } \delta\text{-cover of } S\right\}$$

Which can be used to define the d-dimensional Hausdorff measure $H^d(S) = \lim_{\delta\to 0} H^d_\delta(S)$. Finally, the Hausdorff dimension itself is defined as

$$\dim_{\text{H}}(S) = \inf\{d \geq 0 | H^d(S) = 0\}$$

A discussion of the Hausdorff dimension of the Hilbert curve coordinate functions can be found in [15]. There are many other definitions of fractal dimension which could be used to investigate the graphs of the Hilbert curve's coordinate functions, with each definition having the potential to illuminate a subtly different aspect of the functions' nature.

References

[1] S2 cells. *S2Geometry developer documentation, Google LLC.* `https://s2geometry.io/devguide/s2cell_hierarchy.html` [accessed 05-April-2022].

[2] Michael Bader. *Space-Filling Curves: An Introduction with Applications in Scientific Computing.* Springer-Verlag, Berlin Heidelberg 2013.

[3] Donald R. Bartle, Robert G.; Sherbert. *Introduction to Real Analysis (3rd ed.).* John Wiley and Sons, West Sussex, UK, 2000.

[4] Joseph Warren Dauben. *Introductory functional analysis with applications.* John Wiley and Sons, 1978.

[5] Joseph Warren Dauben. *Georg Cantor: His Mathematics and Philosophy of the Infinite.* Princeton University Press, 1990.

[6] G. A. Edgar. *Measure, Topology, and Fractal Geometry.* Springer-Verlag, New York, 1990.

[7] K. J. Falconer. *Fractal Geometry: Mathematical Foundations and Applications.* John Wiley and Sons, West Sussex, UK, 1990.

[8] K. J. Falconer. *Techniques in Fractal Geometry.* John Wiley and Sons, New York, UK, 1997.

[9] HIROSHI Fukuda, MICHIO Shimizu, and GISAKU Nakamura. New gosper space filling curves. In *Proceedings of the International Conference on Computer Graphics and Imaging (CGIM2001)*, volume 34, page 38, 2001.

[10] Fernando Q Gouvêa. Was cantor surprised? *The American Mathematical Monthly*, 118(3):198–209, 2011.

[11] David Hilbert. Ueber die stetige abbildung einer linie auf ein flächenstück. *Mathematische Annalen*, 38:459–460, 1891.

[12] Stephane Jaffard. Multifractal formalism for functions part i: results valid for all functions. *SIAM Journal on Mathematical Analysis*, 28(4):944–970, 1997.

[13] Stephane Jaffard. Multifractal formalism for functions part ii: self-similar functions. *SIAM Journal on Mathematical Analysis*, 28(4):971–998, 1997.

[14] Nigel Lesmoir-Gordon. *Introducing Fractal Geometry.* Icon Books, 2000), 2000.

[15] Mark McClure. The hausdorff dimension of hilbert's coordinate functions. *Real Analysis Exchange*, pages 875–883, 1998.

[16] E Netto. Beitrag zur mannigfaltigkeitslehre. 1879.

[17] Samuel Nicolay and Laurent Simons. Building cantor's bijection. *arXiv preprint arXiv:1409.1755*, 2014.

[18] Giuseppe Peano. Sur une courbe, qui remplit toute une aire plane. *Mathematische Annalen*, 36(1):157–160, 1890.

[19] Hans Sagan. *Space-Filling Curves.* Springer-Verlag, New York 1994.