Hyperorthogonal well-folded Hilbert curves

Hyperorthogonal Well-Folded Hilbert Curves

Arie Bos and Herman J. Haverkort

Deptartment of Mathematics and Computer Science
Eindhoven University of Technology, The Netherlands
`arie_bos@online.nl`, `cs.herman@haverkort.net`

Abstract

R-trees can be used to store and query sets of point data in two or more dimensions. An easy way to construct and maintain R-trees for two-dimensional points, due to Kamel and Faloutsos, is to keep the points in the order in which they appear along the Hilbert curve. The R-tree will then store bounding boxes of points along contiguous sections of the curve, and the efficiency of the R-tree depends on the size of the bounding boxes—smaller is better. Since there are many different ways to generalize the Hilbert curve to higher dimensions, this raises the question which generalization results in the smallest bounding boxes. Familiar methods, such as the one by Butz, can result in curve sections whose bounding boxes are a factor $\Omega(2^{d/2})$ larger than the volume traversed by that section of the curve. Most of the volume bounded by such bounding boxes would not contain any data points. In this paper we present a new way of generalizing Hilbert's curve to higher dimensions, which results in much tighter bounding boxes: they have at most 4 times the volume of the part of the curve covered, independent of the number of dimensions. Moreover, we prove that a factor 4 is asymptotically optimal.

1998 ACM Subject Classification E.1 Data Structures, F.2.2 Geometrical problems and computations, H.3.1. Indexing methods, H.3.2 File organization

Keywords and phrases space-filling curve, Hilbert curve, multi-dimensional, range query, R-tree

Digital Object Identifier 10.4230/LIPIcs.SOCG.2015.812

1 Introduction

1.1 Space-filling curves and spatial index structures

A d-dimensional space-filling curve is a continuous, surjective mapping from $\mathbb{R}$ to $\mathbb{R}^d$. In the late 19th century Peano [14] described such mappings for $d = 2$ and $d = 3$. Since then, various other space-filling curves have been found, and they have been applied in diverse areas such as spatial databases, load balancing in parallel computing, improving cache utilization in computations on large matrices, finite element methods, image compression, and combinatorial optimization [3, 7, 15]. In this paper we present new space-filling curves for $d > 2$ that have favourable properties for use in spatial data structures.

In particular, we consider data structures for d-dimensional points such as R-trees [12]. In such data structures, data points are organised in blocks, often stored in external memory. Each block contains at most B points, for some parameter B, and each point is stored in exactly one block. For each block we maintain a bounding box, which is the smallest axis-aligned d-dimensional box that contains all points stored in the block. The bounding boxes of the blocks are stored in an index structure, which may often be kept in main memory. To find all points intersecting a given query window Q, we can now query the index structure for all bounding boxes that intersect Q; then we retrieve the corresponding blocks, and check the points in those blocks for answers to our query. We may also use the index structure to find the nearest neighbour to a query point q: if we search blocks in order of increasing

31st International Symposium on Computational Geometry (SoCG'15).
Editors: Lars Arge and János Pach; pp. 812–826

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

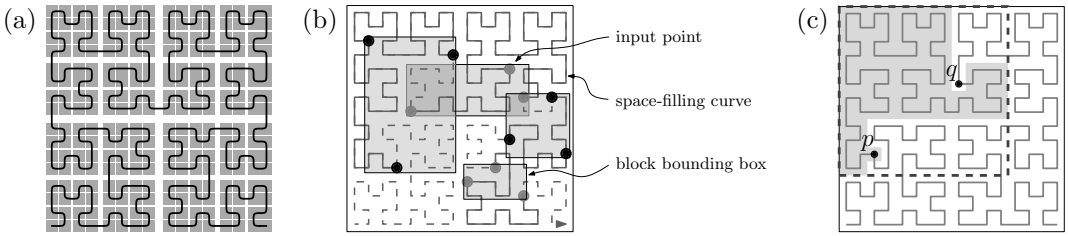

Figure 1 (a) Sketch of Hilbert's space-filling curve. (b) Blocks of an R-tree or similar data structure with $B = 3$. (c) Box-to-curve ratio of the section between p and q = area of the bounding box of the curve section S between p and q, divided by the area covered by S: $12 \cdot 12/87 \approx 1.66$.

distance from q, we will retrieve exactly the blocks whose bounding boxes intersect the largest empty sphere around q. The grouping of points into blocks determines what block bounding boxes are stored in the index structure, and in practice, retrieving these blocks is what determines the query response time [7].

If we store n points in d dimensions with B points in a block, $\Theta((n/B)^{1-1/d})$ blocks may need to be visited in the worst case if the query window is a rectangular box with no points inside [11], and $\Theta(n/B)$ blocks may need to be visited if the query window is an empty sphere. The *Priority-R-tree* achieves these bounds [2], whereas a heuristic solution by Kamel and Faloutsos [10], which is explained below, may result in visiting $\Theta(n/B)$ blocks even if the query window is a rectangular box with no points inside [2]. However, experimental results for (near-)point data and query ranges with few points inside [8] indicate that the approach by Kamel and Faloutsos seems to be more effective in practice for such settings. Moreover, regardless of the type of data and query ranges, a structure based on the ideas of Kamel and Faloutsos is much easier to build and maintain than a Priority-R-tree [2].

Kamel and Faloutsos proposed to determine the grouping of points into blocks as follows: we order the input points along a space-filling curve and then put each next group of B points together in a block (see Figure 1(b)). Note that the number of blocks retrieved to answer a query is simply the number of bounding boxes intersected. Therefore it is important that the ordering induced by the space-filling curve makes us fill each block with points that lie close to each other and thus have a small bounding box.

Kamel and Faloutsos proposed to use the Hilbert curve [9] for this purpose. One way to describe the two-dimensional Hilbert curve is as a recursive construction that maps the unit interval $[0, 1]$ to the unit square $[0, 1]^2$. We subdivide the square into a grid of 2×2 square cells, and simultaneously subdivide the unit interval into four subintervals. Each subinterval is then matched to a cell; thus Hilbert's curve traverses the cells one by one in a particular order. The mapping from unit interval to unit square is refined by applying the procedure recursively to each subinterval-cell pair, so that within each cell, the curve makes a similar traversal. The traversals within these cells are rotated and/or reflected so that the traversal remains continuous from one cell to another (see Figure 1(a)). The result is a fully-specified mapping $f : [0, 1] \to [0, 1]^2$ from the unit interval to the unit square. The mapping is easily reversed, and thanks to the fact that the curve is based on recursive subdivision of a square in quadrants, the reversed mapping can be implemented very efficiently with coordinates represented as binary numbers. This gives us a way to decide which of any two points in the unit square is the first along the curve.

We can sketch the shape of the curve by drawing, for the n-th level of recursion, a polygonal curve, an *approximating curve* A_n, that connects the centres of the 4^n squares in the order in which they are visited. In fact, the mapping f can also be described as

the limit of the approximating curves A_n as n goes to infinity. Explicit descriptions of the approximating curves help us to reason about the shapes of curve sections, and thus, about the extents of their bounding boxes. For ease of notation, in this paper we scale the approximating curve for any level n by a factor 2^n and translate it so that its vertices are exactly the points $\{0, \ldots, 2^n - 1\}^2$.

A d-dimensional version of Hilbert's curve could now be described by a series of curves A_n for increasing n, each visiting the points $\{0, \ldots, 2^n - 1\}^d$. For $d \geq 3$, there are many ways to define such a series of curves [1, 5, 6], but their distinctive properties and their differences in suitability for our purposes are largely unexplored.

1.2 Our results

In this paper we present a family of space-filling curves, for any number of dimensions $d \geq 3$, with two properties which we call *well-foldedness* and *hyperorthogonality*—Hilbert's two-dimensional curve also has these properties. We show that these properties imply that the curves have good *bounding-box quality* as defined by Haverkort and Van Walderveen [7].

More precisely, for any $0 \leq a \leq b \leq 1$, let $f([a, b])$ denote the section of the space-filling curve f from $f(a)$ to $f(b)$, that is, $\bigcup_{a \leq t \leq b} f(t)$. The *box-to-curve ratio (BCR)* of a section $f([a, b])$ is the volume of the minimum axis-aligned bounding box of $f([a, b])$ divided by the volume (d-dimensional Lebesgue measure) of $f([a, b])$, see Figure 1(c). The worst-case BCR of a space-filling curve f is the maximum BCR over all sections of f. We show that the worst-case BCR of a well-folded, hyperorthogonal space-filling curve is at most 4, independent of the number of dimensions. Moreover, we show that this is asymptotically optimal: we prove that any d-dimensional space-filling curve that is described by a series of curves A_n as defined above, has a section with BCR at least $4 - O(1/2^d)$. In contrast, the d-dimensional "Hilbert" curves of Butz [4], as implemented by Moore [13], have sections with BCR in $\Omega(2^{d/2})$.

In Section 1.3 we introduce basic nomenclature and notation. Section 2 defines the concept of well-foldedness, and presents sufficient and necessary conditions for approximating curves of well-folded space-filling curves. Section 3 introduces the concept of hyperorthogonality. We present sufficient and necessary conditions for approximating curves of well-folded space-filling curves to be hyperorthogonal. The necessity of these conditions is then used to prove that any section of a hyperorthogonal well-folded space-filling curve has good box-to-curve ratio. Our next task is to show that hyperorthogonal well-folded curves actually exist, and this is the topic of Section 4. We combine the conditions from the previous sections to learn more about the shape of hyperorthogonal well-folded curves, and in particular about self-similar curves (Section 5). It turns out that in two, three, and four dimensions, there are actually very few self-similar, well-folded, hyperorthogonal curves; in five and more dimensions, more such curves exist. In Section 6, we make a few remarks about how to implement a comparison operator based on self-similar, well-folded, hyperorthogonal curves in any number of dimensions greater than two. Finally, in Section 7, we compare the bounding box quality of hyperorthogonal well-folded curves to lower bounds and to the bounding box quality of Butz's generalization of Hilbert curves, and we discuss directions for further research.

In this extended abstract we omit the proofs of most theorems, lemmas, and observations, as well as many details of the comparison operator discussed in Section 6. We intend to publish the proofs and further details (including pseudocode) of a non-recursive implementation of the comparison operator in a more comprehensive version of this paper. Until that is published, the interested reader is welcome to contact the authors for a version of this abstract that includes an appendix with the proofs and the pseudocode.

1.3 Nomenclature and notation

General notation. By D we denote 2^d. By $\text{sign}(i)$ we denote the sign of i, that is, $\text{sign}(i) = -1$ if $i < 0$; $\text{sign}(i) = 0$ if $i = 0$, and $\text{sign}(i) = 1$ if $i > 0$. By $\text{isneg}(i)$ we denote the function defined by $\text{isneg}(i) = 1$ if $i < 0$, and $\text{isneg}(i) = 0$ if $i \geq 0$.

Vertices, edges, directions and axes. The universe in this article is the integer grid in d dimensions $\mathbb{Z}^d$. A *vertex* is a point $v = (v[1], v[2], \ldots, v[d]) \in \mathbb{Z}^d$. An *edge* e is an ordered pair of vertices (v, w) with distance $||w - v|| = 1$. The *direction* of an edge $e = (v, w)$ is the number $i \in \{-d, \ldots, d\} \setminus \{0\}$ such that $w[|i|] - v[|i|] = \text{sign}(i)$ and $w[j] = v[j]$ if $j \neq |i|$. The *axis* of an edge is the absolute value of its direction. Note that the edges (v, w) and (w, v) have opposite directions, but the same axis. By $\langle e_1, e_2, \ldots \rangle$ we denote a sequence of edges with directions $e_1, e_2, \ldots$.

Curves, length, volume, entry and exit. For the purposes of this paper, a *curve on the grid* is an ordered set of unique vertices where each subsequent pair of vertices forms an edge as defined above. Note that a curve on the grid never visits the same vertex more than once. Henceforth, a *space-filling curve* is always a mapping $f : [0, 1] \to [0, 1]^d$, while any other curve discussed in this paper will be assumed to be a curve on the grid. Since a vertex and a direction determine an edge, a curve can alternatively be described by specifying the starting point and listing the directions of its edges in order. A *free* curve is a curve with a specified shape and orientation but with unspecified location; it is described by the directions of its edges. Note that curves are directed. The *reverse* $\overleftarrow{C}$ of a free curve C is obtained by reversing the order of the edge directions *and* reversing the directions themselves, which means negating them. The *length* of a curve is the number of edges, the *volume* of a subset of the grid is the number of vertices. So the volume $\text{vol}(C)$ of a curve C is its length + 1. The first vertex of a curve is called the *entry*; the last vertex is called the *exit*.

***k*-Curves and *k*-cubes.** A *k-curve* is a Hamiltonian path on the integer grid in a d-dimensional cube with $2^{d \cdot k}$ points, so with a side of length $2^k - 1$. Such a cube is called a *k-cube*. Since each of the (integer) points of the cube is visited by the curve exactly once, the length of a k-curve is $2^{d \cdot k} - 1$ and its volume is $2^{d \cdot k}$.

Approximating curves. The space-filling curves under study in this paper will be approximated by curves on the grid as just defined. By $A_0, A_1, \ldots$ we will denote a sequence of curves that approximates a d-dimensional space-filling curve, where A_0 is a single vertex and A_k is a k-curve. By $v_{k,1}, v_{k,2}, \ldots, v_{k,K}$, where $K = 2^{d \cdot k}$, we denote the vertices of A_k in order, and by $e_{k,i}$ we denote the direction of the edge $(v_{k,i}, v_{k,i+1})$. Recall that each vertex $v_{k,i}$ of A_k represents a d-dimensional hypercube $H_{k,i}$ of width $1/2^k$ that is visited by the space-filling curve approximated by A_k, and the vertices $v_{k+1,D \cdot i - D + 1}, \ldots, v_{k+1,D \cdot i}$ model the order in which the space-filling curve traverses the d-dimensional hypercubes of width $1/2^{k+1}$ whose union is $H_{k,i}$. Therefore it must be possible to construct A_{k+1} from A_k, which we call the *parent curve*, by *inflation*: we replace each vertex $v_{k,i}$ of the parent curve with a 1-curve $C_{k,i}$ (a *child curve*), whose vertices are those of the unit cube, translated by $2 \cdot v_{k,i}$. Each edge $(v_{k,i}, v_{k,i+1})$ of the parent curve is replaced by an edge $(v_{k+1,D \cdot i}, v_{k+1,D \cdot i+1})$ of the same direction, connecting the exit of $C_{k,i}$ to the entry of $C_{k,i+1}$, see Figure 2 (left). Note that not just any choice of child curves results in a valid $(k + 1)$-curve. The 1-curves that replace the vertices have to be chosen carefully such that for each edge $(v_{k,i}, v_{k,i+1})$ of the parent curve, there is indeed an edge in the grid from the exit of $C_{k,i}$ to the entry of

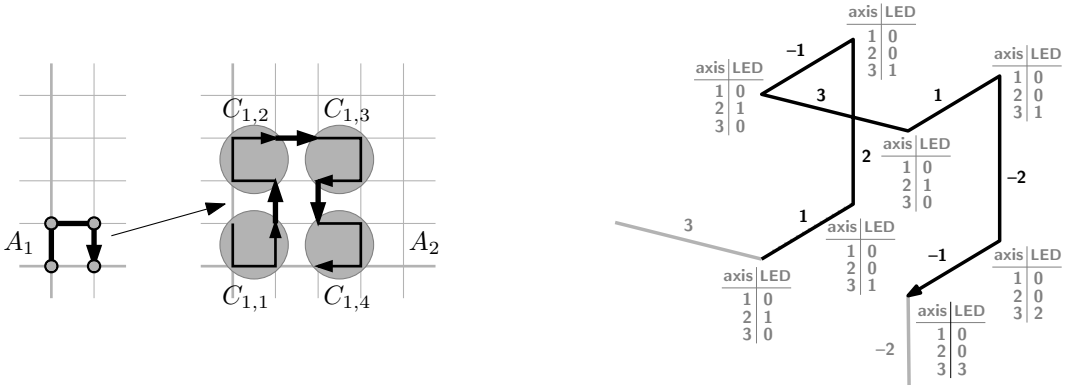

Figure 2 Left: A parent curve A_1 is inflated to create A_2, which is composed of the child curves $C_{1,1}$, $C_{1,2}$, $C_{1,3}$ and $C_{1,4}$, and edges of A_1 which are translated such that they connect the child curves to each other at their end points. Right: $G(3)$ (in black) with the directions of its edges; in grey: $G(3)$ extended with an entry edge $\langle d \rangle$ and an exit edge $\langle -(d-1) \rangle$, with the edge distance table according to Definition 11 for each vertex.

$C_{k,i+1}$. In Section 2 we will discuss how the 1-curves should be constructed so that they match up.

Observe that our definition of curves on the grid restricts the generalizations of Hilbert curves under study to *face-continuous curves*, that is, each pair of consecutive d-dimensional hypercubes along the curve must share a $(d-1)$-dimensional face. In Section 7, we will discuss why, in the context of this paper, this restriction is justified.

2 Well-folded curves

In the process of inflating, we will restrict ourselves in this paper to replacing vertices with isometric images (translations, rotations and reflections) of one particular 1-curve, namely the free curve $G(d)$ that follows the so-called *binary reflected Gray code*:

▶ **Definition 1.** The free curve $G(d)$ is defined recursively as follows: $G(0)$ is empty; $G(d)$ is the concatenation of $G(d-1)$, $\langle d \rangle$, and $\overleftarrow{G}(d-1)$.

For example, $G(2)$ is the free curve $\langle 1, 2, -1 \rangle$, $G(3)$ is shown in Figure 2 (right), and $G(4)$ is the free curve $\langle 1, 2, -1, 3, 1, -2, -1, 4, 1, 2, -1, -3, 1, -2, -1 \rangle$. The length of $G(d)$ is, by induction, $2^d - 1$, which is the maximum length of a Hamiltonian path on the unit cube in $\mathbb{Z}^d$. Notice that in $G(d)$, each edge $\langle a \rangle$ is preceded by an edge $\langle 1 \rangle$ and followed by an edge $\langle -1 \rangle$ if $|a| = 2$, and it is preceded by an edge $\langle -1 \rangle$ and followed by an edge $\langle 1 \rangle$ if $|a| > 2$.

▶ **Definition 2.** A curve is *well-folded* if it is a single vertex, or if it is obtained by inflating a well-folded curve by replacing its vertices by isometric images of $G(d)$. A space-filling curve is well-folded if its approximating curves are well-folded.

Note that in two dimensions, all possible 1-curves are in fact isometric images of $G(2)$, so any face-continuous space-filling curve based on recursive subdivision of a square into four squares must be well-folded (for example, Hilbert's curve or the $\beta\Omega$-curve [17]). In higher dimensions, the most common generalizations of the Hilbert curve are well-folded as well, but there are also face-continuous curves based on recursive subdivision of a cube into eight cubes that are not well-folded (using generators of types B and C from Alber and

Niedermeier [1, 5]). In Section 7, we will briefly get back to non-well-folded curves; until then, we will focus on well-folded curves.

The following lemma will prove useful later:

▶ **Lemma 3.** *The axes of the first (and last) n edges of $G(d)$ constitute the set $\{1,\ldots,m\}$, where $m = 1 + \lfloor \log_2(n) \rfloor = \lceil \log_2(n+1) \rceil$.*

The isometric transformations of 1-curves which we need in this paper are those of the hyperoctahedral group of symmetries of the hypercube. This group is the product of the symmetric group S_d (the group of all permutations of the d coordinate axes) and the group of 2^d reflections formed by all combinations of reflections in hyperplanes orthogonal to the coordinate axes. Thus there are $d! \cdot 2^d$ such transformations.

To distinguish these transformations, we will use *signed permutations*. A signed permutation π working on $\{-d,\ldots,d\} \setminus \{0\}$ is denoted by $[\pi[1], \pi[2], \ldots, \pi[d]]$, where π is the bijection from $\{-d,\ldots,d\} \setminus \{0\}$ to itself defined by $\pi(k) = \pi[k]$ and $\pi(-k) = -\pi[k]$ for $k \in \{1,\ldots,d\}$. Given a k-cube H, a signed permutation π specifies the isometry that maps H onto itself and maps the direction k to the direction $\pi(k)$. If $\pi = [\pi[1], \pi[2], \ldots, \pi[d]]$ is a signed permutation, then $\pi(\mathcal{X})$ denotes the application of π to all elements of the vector, set, or sequence $\mathcal{X}$; $|\pi|$ denotes the permutation $[|\pi_1|, |\pi_2|, \ldots, |\pi_d|]$; and π^{-1} denotes the inverse of π, that is, $\pi^{-1}(x) = y$ if and only if $\pi(y) = x$.

The *orientation* of a 1-curve C, denoted by $\mathrm{or}(C)$, is the direction of the vector from entry to exit. Note that $\mathrm{or}(G(d)) = d$, the direction of the middle edge of $G(d)$. Hence, $\mathrm{or}(\pi(G(d))) = \pi(d)$.

Consider a sequence of well-folded approximating curves $A_0, A_1, \ldots$. Given a particular level k, let K be $2^{d\cdot k}$, and let $\sigma_{k,i}$, for $i \in \{1,\ldots,K\}$, be the transformation (modulo translation) that is applied to $G(d)$ to obtain the 1-curve $C_{k,i}$ that replaces $v_{k,i}$ in the inflation of A_k to A_{k+1}. For example, for the curves in Figure 2 (left) we have $\sigma_{0,1} = [1,2]$; $\sigma_{1,1} = [-1,2]$; $\sigma_{1,2} = [-2,1]$; $\sigma_{1,3} = \sigma_{1,4} = [2,-1]$. As observed before, the 1-curves that replace the vertices have to be chosen carefully such that there is an edge with direction $e_{k,i}$ from the exit of $C_{k,i}$ to the entry of $C_{k,i+1}$. This leads to the following conditions:

▶ **Theorem 4.** *The permutations σ result in a sequence of well-folded approximating curves if and only if, for each k and for each $1 \le i < 2^{d\cdot k}$, we have:*

- *for $j \in \{1,\ldots,d\}$, we have $\mathrm{sign}(\sigma_{k,i}^{-1}(j)) = \mathrm{sign}(\sigma_{k,i+1}^{-1}(j))$ if and only if j equals neither or both of $|\sigma_{k,i}(d)|$ and $|e_{k,i}|$;*
- $\mathrm{sign}(\sigma_{k,i+1}^{-1}(e_{k,i})) = 1$.

Given the edges and the signs of the inverse permutations, Theorem 4 allows us to determine the last elements of each permutation. Conversely, given the edges and the last elements of each permutation, Theorem 4 allows us to determine the signs of each permutation. Note that this leaves $d-1$ elements of each $|\sigma_{k,i}|$ unspecified and without consequence: any permutation of those elements will do.

▶ **Observation 5.** *Let f be a well-folded space-filling curve approximated by $A_0, A_1, \ldots$, and let $x = f(0)$ be the starting point of f. Then $x[j] = \sum_{k=0}^{\infty} \mathrm{isneg}(\sigma_{k,1}^{-1}(j))/2^{k+1}$. In other words, the digits of the binary representation of $x[j]$ behind the fractional point are $\mathrm{isneg}(\sigma_{0,1}^{-1}(j)), \mathrm{isneg}(\sigma_{1,1}^{-1}(j)), \mathrm{isneg}(\sigma_{2,1}^{-1}(j)), \ldots$.*

3 Hyperorthogonal well-folded curves

So far, we have been defining and discussing properties of curves that are in fact common to the best-known previous generalizations of Hilbert's curve to higher dimensions. We will

now introduce a new property that is *not* satisfied by these curves, and will prove useful in designing novel curves with good box-to-curve ratios.

3.1 Definition and characterization

▶ **Definition 6.** We call a curve *hyperorthogonal* if and only if, for any $n \in \{0, \ldots, d-2\}$, each sequence of 2^n consecutive edges have exactly $n+1$ different axes. A space-filling curve is hyperorthogonal if its approximating curves are hyperorthogonal.

Notice that an n-dimensional 1-cube can hold at most $2^n - 1$ consecutive edges of a curve, so any curve constructed by inflation must contain sets of 2^n edges that have at least $n+1$ different axes, for each $n \leq d-1$. Hyperorthogonality requires that this holds for *every* set of 2^n edges, provided[1] $n \leq d-2$. For $d = 2$, hyperorthogonality requires only that each single edge spans a one-dimensional space, which is obvious. So all two-dimensional curves are hyperorthogonal. For $d = 3$ each two consecutive edges must span a two-dimensional space, so each pair of consecutive edges must be orthogonal. (For that reason the property is called 'hyperorthogonal' for higher dimensions as well.) Note that $G(d)$ is hyperorthogonal. As can be seen by inspecting familiar generalizations of Hilbert curves to three dimensions, if we construct a sequence of curves $A_0, \ldots, A_k$ in three or more dimensions by inflation, using isometric images of $G(d)$ to inflate vertices, then A_k is not necessarily hyperorthogonal, even though $G(d)$ is (see, for example, the Butz-Moore curve in Section 5, Figure 3, right, where there are two collinear edges along the top right edge of the cube). The following theorem states what conditions the isometries should fulfill in order to obtain hyperorthogonal curves:

▶ **Definition 7.** The *depth* of a direction a in a signed permutation π, denoted $\mathrm{depth}(\pi, a)$, is defined as follows: if $|a| \in \{|\pi_d|, |\pi_{d-1}|\}$, then $\mathrm{depth}(\pi, a) = 0$, otherwise $\mathrm{depth}(\pi, a)$ is the number j such that $|\pi_{d-1-j}| = |a|$.

▶ **Theorem 8.** *Let K be $2^{d \cdot k}$, and let $A_0, \ldots, A_{k+1}$ be a sequence of well-folded curves constructed by inflation (with all the associated notation introduced in the previous sections). Suppose $A_0, \ldots, A_k$ are hyperorthogonal. Then A_{k+1} is hyperorthogonal as well if and only if the following conditions are satisfied:*

1. *for each $i \in \{1, \ldots, K-1\}$: $\mathrm{depth}(\sigma_{k,i}, e_{k,i}) = \mathrm{depth}(\sigma_{k,i+1}, e_{k,i}) = 0$;*
2. *for each $i \in \{1, \ldots, K-1\}$ and each $a \in \{-d, \ldots, d\} \setminus \{0\}$, we have $|\mathrm{depth}(\sigma_{k,i}, a) - \mathrm{depth}(\sigma_{k,i+1}, a)| \leq 1$.*

3.2 Box-to-curve ratio ≤ 4

To bound the box-to-curve ratio (BCR) of sections of hyperorthogonal well-folded space-filling curves, we will make use of the following lemma:

▶ **Lemma 9.** *For any $n \in \{0, \ldots, d-2\}$, each sequence of 2^n consecutive edges of a well-folded, hyperorthogonal curve lies inside an $(n+1)$-dimensional unit cube.*

▶ **Theorem 10.** *The box-to-curve ratio of any section of a hyperorthogonal well-folded space-filling curve is at most 4.*

[1] The definition leaves little room for being made more strict: raising the bound to $n \leq d-1$ would render hyperorthogonal curves non-existent, at least when $d = 2$.

Table 1 Cases distinguished in the proof of Theorem 10.

Case	*MaxBoxVol*	*MinCrvVol*
A: $2^{d-1}+2 \le \mathrm{vol}(E_k) \le 2^{d+1}$	2^{d+1}	2^{d-1}
B: $2^{d-2}+2 \le \mathrm{vol}(E_k) \le 2^{d-1}+1;\ \mathrm{vol}(Y) \le \mathrm{vol}(X) \le 2^{d-2}$	2^d	2^{d-2}
C: $2^{d-2}+2 \le \mathrm{vol}(E_k) \le 2^{d-1}+1;\ 2^{d-3} < \mathrm{vol}(Y) \le 2^{d-2} < \mathrm{vol}(X)$	$\frac{3}{2}\cdot 2^d$	$\frac{3}{2}\cdot 2^{d-2}$
D: $2^{d-2}+2 \le \mathrm{vol}(E_k) \le 2^{d-1}+1;\ 1 \le \mathrm{vol}(Y) \le 2^{d-3}$	2^d	2^{d-2}
E: $2^{d-2}+2 \le \mathrm{vol}(E_k) \le 2^{d-1}+1;\ \mathrm{vol}(Y) = 0$	2^d	2^{d-2}
F: $3 \le \mathrm{vol}(E_k) \le 2^{d-2}+1$	$4(\mathrm{vol}(E_k)-2)$	$\mathrm{vol}(E_k)-2$
G: $\mathrm{vol}(E_k) \le 2$	2	1

Proof. Consider a section s of a hyperorthogonal well-folded space-filling curve f, approximated by a series of curves $A_0, A_1, \ldots$. Let E_k be the subcurve of A_k that contains all vertices $v_{k,i}$ that represent hypercubes $H_{k,i}$ of width $1/2^k$ that are intersected by s. More specifically, let k be the smallest k such that A_k contains at least one vertex $v_{k,i}$ such that $H_{k,i}$ is fully covered by s. The bounding box of a subcurve $v_{k,h}, \ldots, v_{k,j}$ of A_k is the smallest axis-aligned box that fully contains all hypercubes $H_{k,h}, \ldots, H_{k,j}$.

By our choice of k, E_k contains vertices from at most two (consecutive) child curves $C_{k-1,x}$ and $C_{k-1,y}$ of A_{k-1}, because otherwise one child curve of A_{k-1} would be completely covered by s, contradicting our choice of k. Note that this implies that the bounding box of E_k has at most the volume of two 1-cubes, that is, 2^{d+1}. Without loss of generality, let $C_{k-1,x}$ be the child curve of A_{k-1} that contains the largest part of E_k and call this part X, let Y be the remaining part of E_k (if any), and let $c = |e_{k,\min(x,y)}|$ be the axis of the connecting edge of X and Y. By definition, $\mathrm{vol}(Y) = \mathrm{vol}(E_k) - \mathrm{vol}(X) \le \mathrm{vol}(X)$.

A number of cases with smartly chosen boundaries for $\mathrm{vol}(E_k)$, $\mathrm{vol}(X)$ and $\mathrm{vol}(Y)$ can now be distinguished, as shown in Table 1. In each case, we derive an upper bound *MaxBoxVol* on the bounding box volume, and a lower bound *MinCrvVol* on the number of vertices of E_k that represent hypercubes completely covered by s (this is usually all of E_k except for the first and last vertex). From this we can derive that the box-to-curve ratio is less than $\mathit{MaxBoxVol}/\mathit{MinCrvVol} \le 4$. Note that cases B, C, D, and E are subcases for the same bounds on $\mathrm{vol}(E_k)$, where case B is the case of having small X, and cases C, D, E are the cases of large X with various bounds on the size of Y. For cases A, E, and G the bounds on the bounding box volume are trivial; cases B, C, D, and F require a more careful analysis.

Case B: By Theorem 8, for the axis c of the connecting edge between X and Y we have $\mathrm{depth}(\sigma_{k-1,x}, c) = 0$. Since $\mathrm{vol}(X) \le 2^{d-2}$, Lemma 3 now tells us that the edges of X have axes from $|\sigma_{k-1,x}|(\{1, \ldots, d-2\})$, hence not including c. Therefore X is included in the half of $C_{k-1,x}$ that lies closest to $C_{k-1,y}$. Likewise, Y is included in the half of $C_{k-1,y}$ that lies closest to $C_{k-1,x}$. These two halves together constitute a unit cube of volume 2^d.

Case C: As in case B, Y is included in the half of $C_{k-1,y}$ that lies closest to $C_{k-1,x}$. This half, together with $C_{k-1,x}$, has a bounding box of volume $\frac{3}{2}\cdot 2^d$. The minimum curve volume *MinCrvVol* is at least $\mathrm{vol}(E_k) - 2 = \mathrm{vol}(X) + \mathrm{vol}(Y) - 2 \ge 2^{d-2} + 2^{d-3} = \frac{3}{2}\cdot 2^{d-2}$.
Case D: Given the bounds on $\mathrm{vol}(X)$ and $\mathrm{vol}(Y)$, Lemma 3 tells us that the edges of X have axes from $|\sigma_{k-1,x}|(\{1, \ldots, d-1\})$, and the edges of Y have axes from $|\sigma_{k-1,y}|(\{1, \ldots, d-3\})$. Now let a be $|\sigma_{k-1,x}(d)|$. By Theorem 8, $\mathrm{depth}(\sigma_{k-1,y}, a) \le \mathrm{depth}(\sigma_{k-1,x}, a) + 1 = 1$ and therefore a is not included in $|\sigma_{k-1,y}|(\{1, \ldots, d-3\})$. If $a = c$, it follows that X

and Y lie in half-cubes that together constitute a unit cube of volume 2^d, as in case B. Otherwise, if $a \neq c$, it follows that E_k may contain multiple edges of direction c but does not include any edge with direction a. Therefore E_k lies completely in a box that spans two 1-cubes in dimension c, half a 1-cube in dimension a, and one 1-cube in the remaining dimensions. The volume of this box is 2^d.

Case F: By Lemma 9, each set of $\mathrm{vol}(E_k) - 1$ edges of A_k is contained in a unit cube of $\lceil \log(\mathrm{vol}(E_k)-1) \rceil + 1 = \lfloor \log(\mathrm{vol}(E_k)-2) \rfloor + 2$ dimensions, of volume at most $4(\mathrm{vol}(E_k)-2)$. ◀

4 General construction method

In Section 2, Theorem 4, we learned about sufficient and necessary conditions for well-folded curves in general, and in Section 3, Theorem 8, we learned about specific conditions for hyperorthogonal well-folded curves. It remains to show that curves satisfying both the general and the specific conditions actually exist. In this section we will combine the conditions of Theorems 4 and 8 to derive conditions on the entry and exit points and isometries used in the construction of hyperorthogonal well-folded curves. We will show how to construct curves that satisfy all conditions, for any $d \geq 3$ (recall that for $d = 2$, we have Hilbert's curve).

▶ **Definition 11.** The *edge distance* of the axis a to the vertex v within the curve C, denoted $\mathrm{dist}(C, v, a)$, is the distance along C between v and the closest edge with axis a; more precisely, $\mathrm{dist}(C, v, a)$ is one less than the length of the smallest subcurve of C that includes v and an edge with axis a. (For a small example, see Figure 2, right.)

Theorem 8 has a remarkable consequence:

▶ **Lemma 12.** *In well-folded hyperorthogonal curves,* $\mathrm{depth}(\sigma_{k,i}, a) \leq \mathrm{dist}(A_k, v_{k,i}, a)$.

Lemma 12 gives us the following idea for an algorithm to specify the permutations $|\sigma_{k,i}|$, except for the order of the last two elements: simply sort all axes a by order of decreasing edge distance $\mathrm{dist}(A_k, v_{k,i}, a)$. In fact, as we will show now, a version of this algorithm that only considers edge distances within small subcurves suffices. We choose an *entry direction* and an *exit direction* and denote these by $e_{k,0} = e_{0,0}$ and $e_{k,K} = e_{0,1}$, respectively, for any k and $K = 2^{d \cdot k}$.

▶ **Definition 13.** Define the *extended child curve* $C'_{k-1,j}$ as the concatenation of an edge $\langle e_{k-1,j-1} \rangle$, the curve $C_{k-1,j}$, and an edge $\langle e_{k-1,j} \rangle$. We define the *local* edge distance $\mathrm{ldist}_{k,i}(a)$ as $\mathrm{dist}(C'_{k-1,j}, v_{k,i}, a)$, where $j = \lceil i/D \rceil$ and $C_{k-1,j}$ is the child curve that contains $v_{k,i}$.

▶ **Lemma 14.** *Suppose that, for* $k \in \{1, 2, \ldots\}$*, we construct the permutations* $\sigma_{k,i}$ *of a well-folded space-filling curve such that the elements of* $|\sigma_{k,i}|$ *are sorted by order of decreasing local edge distance* $\mathrm{ldist}_{k,i}$*. Then each curve* A_k *satisfies the conditions of Theorem 8.*

The above lemma still leaves the order of the last two elements of each $|\sigma_{k,i}|$ undetermined, since these are always the two axes with edge distance zero. To prove that hyperorthogonal well-folded curves exist, it now suffices to show that we can order the last two elements and choose the signs of each $\sigma_{k,i}$ such that the conditions of Theorem 4 are satisfied. We obtain:

▶ **Theorem 15.** *For each choice of* $e_{0,0}$ *and* $e_{0,1}$ *and for each choice for the signs of* $\sigma_{k,1}^{-1}(j)$ *for all* k *and* j*, satisfying* $\mathrm{sign}(\sigma_{k,1}^{-1}(e_{k,0})) = 1$ *for all* k*, there is a unique hyperorthogonal, well-folded space-filling curve* f *approximated by* $A_0, A_1, \ldots$ *in which the elements of each permutation* $|\sigma_{k,i}|$ *are sorted by order of decreasing local edge distance* $\mathrm{ldist}_{k,i}$*.*

Proof. For each level k, we generate A_k as follows. We loop over all $i \in \{1, \ldots, K-1\}$, where $K = 2^{d \cdot k}$, and proceed as follows. The conditions of Theorem 4 require $\text{sign}(\sigma^{-1}_{k,i+1}(e_{k,i})) = 1$. We now choose $|\sigma_{k,i}(d)|$ such that $|\sigma_{k,i}(d)| = |e_{k,i}|$ if and only if $\text{sign}(\sigma^{-1}_{k,i}(e_{k,i})) = 1$: this is always possible since $|e_{k,i}|$ is among the last two elements of $|\sigma_{k,i}|$ whose order was undetermined. Thus we satisfy the first condition of Theorem 4 for $j = |e_{k,i}|$. With $|\sigma_{k,i}|$ completely determined, we can now fill in the remaining signs of $\sigma_{k,i+1}$ such that they fulfill the first condition of Theorem 4. Finally, we determine $|\sigma_{k,K}(d)|$ as dictated by the exit direction $e_{k,K}$ in the same way as we determined $|\sigma_{k,i}(d)|$ for $i < K$. ◀

5 Self-similar curves

By Observation 5, a choice of signs of $\sigma^{-1}_{k,1}(j)$ for all k and j specifies the starting point $f(0)$ of the space-filling curve f in Theorem 15. Thus, the proof of Theorem 15 is a constructive proof that a hyperorthogonal, well-folded space-filling curve exists for any choice of starting point on the boundary of the unit hypercube.

In a practical setting, such as described in Section 1.1, one may want to sort points in the order in which they appear along the curve. To this end we need a comparison operator that decides which of any two given points p and q comes first along the curve. We can do so by determining the largest k such that there is a hypercube $H_{k,i}$, corresponding to a vertex $v_{k,i}$, which contains both points. Then we can use $\sigma_{k,i}$ to determine in which order the 2^d subcubes of this hypercube are traversed, and in particular, in which order this traversal visits the two subcubes containing p and q. The efficiency of the comparison operator now depends on how efficiently we can determine $\sigma_{k,i}$ for any k and i. Unfortunately, straightforward application of Theorem 15 would require us to traverse all of A_k from $v_{k,1}$ to $v_{k,i}$ to determine $\sigma_{k,i}$.

To enable us to determine $\sigma_{k,i}$ more efficiently, we will, in this section, restrict the curves to be *self-similar*, that is, A_{k+1} is the concatenation of 2^d isometric and/or reversed copies of A_k. We will analyse how the choice of the entry of $C_{1,1}$ propagates to the other child curves of A_1, and derive conditions that starting points of self-similar, hyperorthogonal, well-folded space-filling curves should fulfill. It turns out that for any $d \geq 3$, only two different starting points (modulo rotation and reflection) exist for such curves.

For the purposes of this section, the following notation will be helpful.

▶ **Definition 16.** The relative coordinate vector of a vertex v is the vector r such that $r[j] = 0$ if $x[j] \bmod 4 \in \{0, 3\}$, and $r[j] = 1$ if $x[j] \bmod 4 \in \{1, 2\}$.

Note that the relative coordinates of a vertex $v_{k+1,i}$ tell us, for each dimension, whether the vertex is on the outside (0) or on the inside (1) with respect to the 2-cube of A_{k+1} corresponding to the vertex $v_{k-1,j}$ of A_{k-1}, where $j = \lfloor i/D^2 \rfloor$.

Let $\text{ent}_{k,i}, \text{ext}_{k,i} : \{1, \ldots, d\} \to \{0, \ldots, 2^{k+1} - 1\}$ be functions that give the coordinates of the entry and exit point of $C_{k,i}$, that is, the entry point of $C_{k,i}$ has coordinates $(\text{ent}_{k,i}(1), \ldots, \text{ent}_{k,i}(d))$ and the exit point has coordinates $(\text{ext}_{k,i}(1), \ldots, \text{ext}_{k,i}(d))$. Note that $\text{ent}_{k,i}(j) = \text{isneg}(\sigma^{-1}_{k,i}(j)) \pmod 2$, and $\text{ext}_{k,i}(j) = \text{ent}_{k,i}(j) \pmod 2$ if and only if $|\sigma^{-1}_{k,i}(j)| \neq d$. Similarly, let $\text{rlent}_{k,i}, \text{rlext}_{k,i} : \{1, \ldots, d\} \to \{0, 1\}$ be functions that give us the *relative* coordinates of the entry and exit point of $C_{k,i}$. Note that we have $\text{rlent}_{k,i}(j) = (\text{ent}_{k,i}(j) + v_{k,i}[j]) \bmod 2$, and $\text{rlext}_{k,i}(j) = (\text{ext}_{k,i}(j) + v_{k,i}) \bmod 2$. Observe that if $\text{rlent}_{k,i}$ and $v_{k,i}$ are given, this determines $\text{ent}_{k,i}$ and hence, the signs of $\sigma^{-1}_{k,i}$.

▶ **Lemma 17.** *If $A_0, A_1, \ldots$ approximate a self-similar, well-folded, hyperorthogonal space-filling curve f, then each extended child curve $C'_{k,i}$, according to Definition 13, is an isometry of either:*

- *the concatenation of* $\langle d\rangle$, $G(d)$, *and* $\langle -(d-1)\rangle$ *(henceforth called type 0);*
- *the concatenation of* $\langle d-1\rangle$, $G(d)$, *and* $\langle d\rangle$ *(henceforth called type 1).*

We will denote the type of the child curve $C_{k,i}$ by $T_{k,i}$. A direct consequence of Lemma 17 is that we may assume, without loss of generality (modulo reflection, rotation and reversal), that $C_{0,1} = A_1 = G(d)$ with entry direction d and exit direction $-(d-1)$, with $T_{0,1} = 0$. Moreover, we should have $v_{2,1}[d] = 0$ and $v_{2,K}[d-1] = 0$, where $K = D^2 = (2^d)^2$, so that the child curves $C_{1,1}$ and $C_{1,D}$ can be extended with, respectively, the same entry edge $\langle d\rangle$ and the same exit edge $\langle -(d-1)\rangle$ as $C_{0,1}$. By tracing the relative coordinates of the entry and exit points through the child curves of A_1, using the conditions of Theorems 4 and 8, we now find the following:

▶ **Lemma 18.** $\mathrm{rlext}_{1,D} = \mathrm{rlent}_{1,1} \circ \omega$, *where* $\omega = [d-1, 2, \ldots, d-2, d, 1]$.

When we inflate A_2 to obtain A_3, so that a 2-curve replaces each vertex of A_1, the relative coordinates of each 2-curve's exit point should equal the relative coordinates of the next 2-curve's entry point. Because of self-similarity, the 2-curve replacing $v_{1,i}$ must itself be an isometry of either A_2 (if $T_{1,i} = 0$) or $\overleftarrow{A}_2$ (if $T_{1,i} = 1$). As a result of the transformation $\sigma_{1,i-1}$, the relative coordinates of the exit point of the 2-curve replacing $v_{1,i-1}$ are given by the function $\mathrm{rlent}_{1,1} \circ \omega \circ |\sigma_{1,i-1}^{-1}|$ if $T_{1,i-1} = 0$, and by $\mathrm{rlent}_{1,1} \circ |\sigma_{1,i-1}^{-1}|$ if $T_{1,i-1} = 1$. The relative coordinates of the entry point of the 2-curve replacing $v_{1,i}$ are given by the function $\mathrm{rlent}_{1,1} \circ |\sigma_{1,i}^{-1}|$ if $T_{1,i} = 0$, and by $\mathrm{rlent}_{1,1} \circ \omega \circ |\sigma_{1,i}^{-1}|$ if $T_{1,i} = 1$. Thus we get:

▶ **Lemma 19.**
If $T_{1,i-1} = 0$ *and* $T_{1,i} = 0$, *we have* $\mathrm{rlent}_{1,1} \circ \omega \circ |\sigma_{1,i-1}^{-1}| = \mathrm{rlent}_{1,1} \circ |\sigma_{1,i}^{-1}|$.
If $T_{1,i-1} = 0$ *and* $T_{1,i} = 1$, *we have* $\mathrm{rlent}_{1,1} \circ \omega \circ |\sigma_{1,i-1}^{-1}| = \mathrm{rlent}_{1,1} \circ \omega \circ |\sigma_{1,i}^{-1}|$.
If $T_{1,i-1} = 1$ *and* $T_{1,i} = 0$, *we have* $\mathrm{rlent}_{1,1} \circ |\sigma_{1,i-1}^{-1}| = \mathrm{rlent}_{1,1} \circ |\sigma_{1,i}^{-1}|$.
If $T_{1,i-1} = 1$ *and* $T_{1,i} = 1$, *we have* $\mathrm{rlent}_{1,1} \circ |\sigma_{1,i-1}^{-1}| = \mathrm{rlent}_{1,1} \circ \omega \circ |\sigma_{1,i}^{-1}|$.

We can now analyse the possible successions of types $T_{1,i}$ and permutations $\sigma_{1,i}$ for $i \in \{1, \ldots, 2^d\}$ and prove that Lemma 19 can only be true if:

▶ **Lemma 20.** $\mathrm{rlent}_{1,1}(j) = \mathrm{rlent}_{1,1}(j-1)$ *for all* $j \in \{2, \ldots, d-1\}$.

By exploiting self-similarity recursively, we now find:

▶ **Lemma 21.** $\mathrm{rlent}_{k,1} = \mathrm{rlent}_{1,1}$ *for all* $k \geq 1$.

This leads almost directly to:

▶ **Theorem 22.** *If* f *is a hyperorthogonal well-folded space-filling curve mapping* $[0,1]$ *to* $[0,1]^d$, *then, modulo reflection, reversal and rotation,* $f(0)$ *is either* $(0, \ldots, 0, 0)$ *or* $(\frac{1}{3}, \ldots, \frac{1}{3}, 0)$.

In fact, such curves exist for any $d \geq 3$:

▶ **Theorem 23.** *For any* $d \geq 3$, *there are self-similar, hyperorthogonal, well-folded* d*-dimensional space-filling curves starting at* $(0, \ldots, 0, 0)$ *and* $(\frac{1}{3}, \ldots, \frac{1}{3}, 0)$.

It turns out that there are actually very few such curves for $d = 3$ and $d = 4$:

▶ **Observation 24.** *If* $d = 3$ *or* $d = 4$, *Lemma 12 leaves no choice with respect to the last two elements, the third-last element, and the first element of the permutations* $|\sigma_{k,i}|$ *in a self-similar curve.*

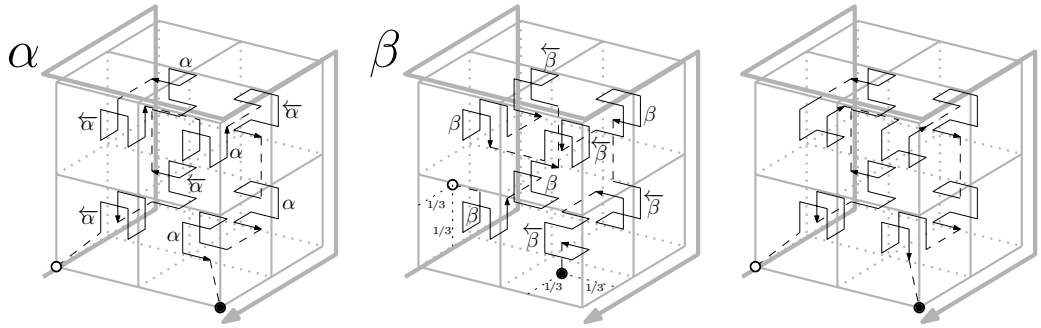

Figure 3 The three-dimensional, self-similar, hyperorthogonal, well-folded space-filling curve with starting points $(0,0,0)$ (left) and $(\frac{1}{3},\frac{1}{3},0)$ (centre), and the three-dimensional curve by Butz and Moore (right). The bold grey curve shows A_1. The solid black curves depict the child curves of A_1, the dashed lines between them indicate how they are connected. The symbols next to the child curves indicate whether they are type 0 (without arrow), or its reverse, type 1 (with arrow). For the Butz-Moore curve, no such indications are given, because the curve is symmetric and there is no need to distinguish between reflections and reversals. The white and black dots on the outer cube indicate the location of $f(0)$ and $f(1)$.

▶ **Corollary 25.** *If $d = 3$ or $d = 4$, there are exactly two self-similar, hyperorthogonal, well-folded d-dimensional space-filling curves.*

Proof. For self-similar curves, we may assume the entry and exit direction to be fixed at $\langle d\rangle$ and $\langle -(d-1)\rangle$, respectively. For the starting point, that is, the signs of $\sigma_{k,1}^{-1}(j)$ for all k and j, only two combinations are possible (Theorem 22). Theorem 15 states that this leads to two unique hyperorthogonal, well-folded space-filling curves in which the elements of each $|\sigma_{k,i}|$ are sorted by order of decreasing local edge distance $\mathrm{ldist}_{k,i}$. By Observation 24, for $d = 3$ and $d = 4$, there is no other way to order the elements of each $|\sigma_{k,i}|$. ◀

The two three-dimensional self-similar, hyperorthogonal, well-folded space-filling curves are illustrated in Figure 3, left and centre.

6 Implementation in software

It is relatively easy to implement an efficient comparison operator that decides which of any two given points comes first along a d-dimensional self-similar hyperorthogonal well-folded space-filling curve. For a fixed choice of space-filling curve f, a recursive implementation would take as input two points $p, q \in [0,1)^d$ that need to be compared, along with a signed permutation σ that specifies how the given curve is placed in the unit cube, and the direction of the curve (forward or reversed). Let $S(p)$ and $S(q)$ be the subcubes of width $1/2$ that contain p and q, respectively.

If $p = q$, one point does not precede the other. Otherwise, if $S(p) \neq S(q)$, one can decide immediately which point comes first, based on the relative order of the vertices that represent $S(p)$ and $S(q)$ along the approximating 1-curve $\sigma(G(d))$. Finally, if $S(p) = S(q)$, that is, p and q lie in the same subcube of width $1/2$, then their relative order can be decided by a recursive call with:

- the points p and q, scaled and translated according to the transformation that maps $S(p)$ to the unit cube;
- the signed permutation and direction that specifies how the space-filling curve traverses $S(p)$.

In fact, thanks to the structure of the approximating curve $\sigma(G(d))$, one can examine the coordinates of p and q one by one, from the coordinate in dimension $|\sigma|(d)$ to the coordinate in dimension $|\sigma|(1)$: as soon as a coordinate is found in which the binary representations of the fractional parts of p and q differ in the first bit, one can decide which of the two points precedes the other. Only if p and q are equal in the first bits of all coordinates, the algorithm needs to go in recursion.

To be able to make the recursive call, the algorithm needs to determine the permutation to use in recursion, that is, the transformation that maps the complete space-filling curve f to the section within $S(p)$, modulo scaling and translation. For the curves described by the constructions of Lemma 14 and Theorem 23 this is relatively straightforward. To determine the unsigned permutation to be used in recursion, we sort the d coordinate axes by decreasing local edge distance from $S(p)$. This sorted list of axes can be constructed on the fly in $\Theta(d)$ time while examining the d coordinates of p and q to decide in which subcube they lie. By Lemma 14, the sorted list of axes gives us the (unsigned) permutation to use in recursion. The signs of the permutation to use in recursion now follow from applying the observations on relative entry points and permutation signs calculated in the previous section.

If the binary representations of the coordinates of p and q consist of k bits per coordinate, then the complete comparison operator runs in $O(d \cdot k)$ time.

7 Evaluation

7.1 Comparing to the Butz-Moore curves

The generalization of Hilbert's curve to d dimensions by Butz [4], as implemented by Moore [13], is a self-similar well-folded curve with starting point in the origin, in which the orientations (and therefore, the signs of the inverse permutations) of the child curves of A_1 are the same as in our hyperorthogonal well-folded curve. Concretely, $|\sigma_{1,i}[d]| = 1$ for $i \in \{1, 2^d\}$, and $|\sigma_{1,i}[d]| = \max(|e_{1,i-1}|, |e_{1,i}|)$ for $1 < i < 2^d$. However, otherwise the permutations are different: all permutations in the Butz-Moore curves are rotations (in the permutation sense of the word), so $|\sigma_{1,i}[j]| = |\sigma_{1,i}[d]| + j \pmod d$. For a graphical description of the 3-dimensional curve, see Figure 3 (right).

Assuming $d \geq 3$, the curve $G(d)$ contains a sequence $\langle 1, 2, -1, (2 + \lfloor d/2 \rfloor), 1\rangle$ or $\langle 1, -2, -1, (2 + \lfloor d/2 \rfloor), 1\rangle$, so there is an i such that $|\sigma_{1,i}(d)| = 2$, $|e_{1,i}| = 1$, and $|\sigma_{1,i+1}(d)| = 2 + \lfloor d/2 \rfloor$. We can now calculate that the curve through the last $2^{\lfloor d/2 \rfloor - 1} + 1$ vertices of $C_{1,i}$ and the first $2^{\lfloor d/2 \rfloor - 1} + 1$ vertices of $C_{1,i+1}$ has box-to-curve ratio at least $2^{d-1}/(2^{d/2} + 2)$, and thus:

▶ **Theorem 26.** *The Butz-Moore curve contains sections with box-to-curve ratio* $\Omega(2^{d/2})$.

The worst-case box-to-curve ratio of the Butz-Moore curves is thus in sharp contrast with the worst-case box-to-curve ratio of our hyperorthogonal, well-folded curves, which have BCR at most 4 for any d. For verification we also calculated the actual worst-case BCR values for $d \in \{2, 3, 4, 5, 6\}$ with the software from Sasburg [16] (Table 2). Further investigations may be done into average BCR values over curve sections of a given size, both for the hyperorthogonal and the Butz curves.

It should be noted, however, that BCR may not be the only relevant measure of bounding-box quality. Haverkort and Van Walderveen [7] argued that, at least for $d = 2$, the size of the *boundary* of a bounding box may be as important as its volume—although volume and boundary size are usually correlated. Using Sasburg's software with a generalization of the worst-case bounding box perimeter ratio from Haverkort and Van Walderveen to higher

Table 2 Worst-case box-to-curve ratios for various curves in up to 6 dimensions.

curve	$d=2$	$=3$	$=4$	$=5$	$=6$	≥ 7
lower bound face-continuous	*2.00*	*2.54*	*3.15*	*3.54*	*3.76*	*4−16/(2^d+3)*
best claimed non-self-sim.	2.22[a]	2.89[b]				
self-sim. hyperorth. well-fdd. $f(0)=(0,\ldots,0,0)$	2.40[c]	3.11	3.53	3.76	3.88	≤ 4
self-sim. hyperorth. well-fdd. $f(0)=(\frac{1}{3},\ldots,\frac{1}{3},0)$		3.14	3.67	3.83	3.92	≤ 4
lower bound non-face-continuous	*3.00*	*3.50*	*3.75*	*3.87*	*3.93*	*4−4/2^d*
Butz-Moore	2.40[c]	3.11	4.74	7.08	10.65	$\Omega(2^{d/2})$

[a] $\beta\Omega$-curve [17] analysed by H&vW [7]; [b] Iupiter [5]; [c] Hilbert's curve [9]

dimensions, we found that already for $d=3$, the self-similar hyperorthogonal well-folded curve with starting point $(\frac{1}{3},\frac{1}{3},0)$ outperforms the Butz curve.

7.2 What about other curves?

In this work we study space-filling curves that can be described by a series of approximating curves $A_0, A_1, \ldots, A_n$, where A_k is a curve on the k-cube. Within this context, we restricted our search for curves with good worst-case BCR first to face-continuous curves; then, more specifically, to well-folded curves; then to hyperorthogonal well-folded curves; and finally to self-similar, hyperorthogonal, well-folded curves. We found that if $d=3$ or $d=4$, there are only two self-similar hyperorthogonal well-folded space-filling curves. For $d=5$ and up, there are many more, as Lemma 12 then starts to leave room for swaps among the first elements of the permutations $\sigma_{k,i}$. We will now address the question of how much room for further improvement there is within these restrictions or if some of these restrictions are dropped.

For $d=2$, Haverkort and Van Walderveen [7] report that the BCR of any section of the well-folded, non-self-similar $\beta\Omega$-curve [17] is 2.22 in the worst case, and for $d=3$, Haverkort [5] claims a fairly complicated, non-self-similar, face-continuous curve with a worst-case BCR of 2.89. These constructions, which do not easily generalize to higher dimensions, constitute improvements of less than 10% with respect to the self-similar hyperorthogonal well-folded curves. For larger values of d, no face-continuous curve can be much better than any hyperorthogonal well-folded curve, since the first is subject to a lower bound that quickly approaches the upper bound of the latter as d grows:

▶ **Theorem 27.** *If f is a space-filling curve approximated by a series of curves $A_0,\ldots,A_k$ within the framework of Section 1.3, then f has a section with BCR at least $4-16/(2^d+3)$.*

The proof is based on the fact that any such curve must contain a sequence of at most $2^{d-2}+1$ edges that have all axes $\{1,\ldots,d\}$. For the specific case of $d=2$, Haverkort and Van Walderveen [7] prove a stronger lower bound of 2.

Now suppose we drop the restriction to face-continuous curves. More precisely, suppose we have a space-filling curve approximated by a sequence of curves on the grid $A_0, A_1, \ldots$, where we allow our curves on the grid to have diagonal edges, that is, we allow any edge (v,w) such that $w\neq v$ and $|w[j]-v[j]|\leq 1$ for all $j\in\{1,\ldots,d\}$. In that case, the lower bound becomes even worse:

▶ **Theorem 28.** *If there is a k and i such that $v_{k,i}$ and $v_{k,i+1}$ differ in at least two coordinates (in other words: if there is a diagonal edge), then f has a section with BCR at least $4-4/2^d$.*

Note that, as Table 2 shows, at least for d up to 6 the lower bound of Theorem 28 for curves with "diagonal edges" is greater than the worst-case BCR of the best hyperorthogonal,

well-folded curves, and for higher dimensions the difference between the lower bound and the upper bound is less than 1%. Therefore, in terms of worst-case BCR, little is to be expected from non-face-continuous curves based on inflation of k-cubes for increasing k.

The question remains whether there are hyperorthogonal curves that are not well-folded, and if so, whether such curves would also have good bounds on the box-to-curve ratio. In other words: is well-foldedness really required in Theorem 10? However, Theorem 27 shows that in any case, there is not much room for finding curves with a better worst-case BCR.

References

1 J. Alber and R. Niedermeier. On multidimensional curves with Hilbert property. *Theory of Computing Systems*, 33(4):295–312, 2000.

2 L. Arge, M. de Berg, H. Haverkort, and K. Yi. The Priority R-tree: a practically efficient and worst-case optimal R-tree. *ACM Tr. Algorithms*, 4(1):9, 2008.

3 M. Bader. *Space-filling curves: an introduction with applications in scientific computing.* Springer, 2013.

4 A. R. Butz. Alternative algorithm for Hilbert's space-filling curve. *IEEE Trans. Comp.*, 20(4):424–426, 1971.

5 H. Haverkort. An inventory of three-dimensional Hilbert space-filling curves. *CoRR*, abs/1109.2323, 2011.

6 H. Haverkort. Harmonious Hilbert curves and other extradimensional space-filling curves. *CoRR*, abs/1211.0175, 2012.

7 H. Haverkort and F. van Walderveen. Locality and bounding-box quality of two-dimensional space-filling curves. *Computational Geometry*, 43(2):131–147, 2010.

8 H. Haverkort and F. van Walderveen. Four-dimensional Hilbert curves for R-trees. *ACM J. Experimental Algorithmics*, 16:3.4, 2011.

9 D. Hilbert. Über die stetige Abbildung einer Linie auf ein Flächenstück. *Math. Ann.*, 38(3):459–460, 1891.

10 I. Kamel and C. Faloutsos. On packing R-trees. In *Conf. on Information and Knowledge Management*, pages 490–499, 1993.

11 K. V. R. Kanth and A. K. Singh. Optimal dynamic range searching in non-replicating index structures. In *Int. Conf. Database Theory, LNCS 154*, pages 257–276, 1999.

12 Y. Manolopoulos, A. Nanopoulos, A. N. Papadopoulos, and Y. Theodoridis. *R-trees: Theory and Applications.* Springer, 2005.

13 D. Moore. Fast Hilbert curve generation, sorting, and range queries. `http://web.archive.org/web/www.caam.rice.edu/~dougm/twiddle/Hilbert/`, 2000, retrieved 23 March 2015.

14 G. Peano. Sur une courbe, qui remplit toute une aire plane. *Math. Ann.*, 36(1):157–160, 1890.

15 H. Sagan. *Space-Filling Curves.* Universitext. Springer, 1994.

16 S. Sasburg. Approximating average and worst-case quality measure values for d-dimensional space-filling curves. Master's thesis, Eindhoven University of Technology, 2011.

17 J.-M. Wierum. Definition of a new circular space-filling curve: $\beta\Omega$-indexing. Technical Report TR-001-02, Paderborn Center for Parallel Computing PC^2, 2002.