Discrete Mathematics and Theoretical Computer Science **AC**, 2003, 53–68

Approximation and Analytical Studies of Inter-clustering Performances of Space-Filling Curves

Ho-Kwok Dai[1] and Hung-Chi Su[2]

[1] *Computer Science Department, Oklahoma State University, Stillwater, Oklahoma 74078, U. S. A.*
[2] *Department of Computer Science, Arkansas State University, State University, Arkansas 72467, U. S. A.*
`dai@cs.okstate.edu, suh@csm.astate.edu`

A discrete space-filling curve provides a linear traversal/indexing of a multi-dimensional grid space. This paper presents an application of random walk to the study of inter-clustering of space-filling curves and an analytical study on the inter-clustering performances of 2-dimensional Hilbert and z-order curve families. Two underlying measures are employed: the mean inter-cluster distance over all inter-cluster gaps and the mean total inter-cluster distance over all subgrids. We show how approximating the mean inter-cluster distance statistics of continuous multi-dimensional space-filling curves fits into the formalism of random walk, and derive the exact formulas for the two statistics for both curve families. The excellent agreement in the approximate and true mean inter-cluster distance statistics suggests that the random walk may furnish an effective model to develop approximations to clustering and locality statistics for space-filling curves. Based upon the analytical results, the asymptotic comparisons indicate that z-order curve family performs better than Hilbert curve family with respect to both statistics.

Keywords: space-filling curves, Hilbert curves, z-order curves, clustering, random walk

---

1 Preliminaries

The subject of space-filling curves has fascinated mathematicians since late 19th century, and has many applications in algorithms, databases, and parallel computation, in which linearization techniques of multi-dimensional arrays or grids are needed. Sample applications include heuristics for Hamiltonian traversals, multi-dimensional space-filling indexing methods [BBK01], image compression, and dynamic unstructured mesh partitioning. For a comprehensive historical development of classical space-filling curves, see [Sag94].

For positive integer n, denote $[n] = \{1, 2, \ldots, n\}$. An m-dimensional (discrete) space-filling curve of length n^m is a bijective mapping $C : [n^m] \to [n]^m$, thus providing a linear indexing/traversal or total ordering of the grid points in $[n]^m$. An m-dimensional grid is said to be of order k if it has side-length $n = 2^k$; a space-filling curve has order k if its codomain is a grid of order k. An m-dimensional space-filling curve C is continuous if the Euclidean distance between $C(i)$ and $C(i+1)$ is 1 for all $i \in [n^m - 1]$. The generation of a sequence of multi-dimensional space-filling curves of successive orders usually follows a recursive framework (on the dimensionality and order), which results in a few classical families, such as Gray-coded curves, Hilbert curves, Peano curves, and z-order curves (see, for examples, [AN00] and [MJFS01]).

Denote by H_k^m and Z_k^m an m-dimensional Hilbert and z-order, respectively, space-filling curve of order k. Figure 1 illustrates the recursive constructions of H_k^2 and Z_k^2 for $m = 2$, and $k = 1, 2$.

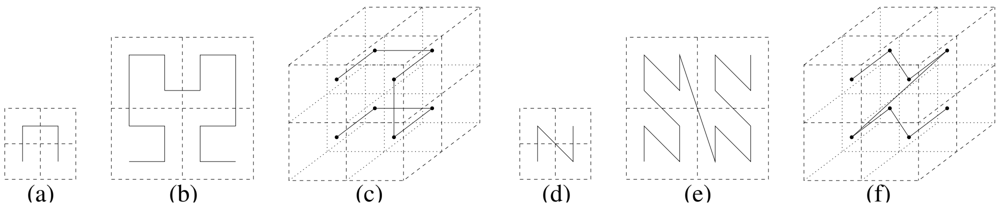

Fig. 1: Recursive constructions of Hilbert and z-order curves of higher order (H_k^m and Z_k^m, respectively) by interconnecting symmetric (via reflection and rotation) subcurves of lower order (H_{k-1}^m and Z_{k-1}^m, respectively): (a) H_1^2; (b) H_2^2; (c) H_1^3; (d) Z_1^2; (e) Z_2^2; (f) Z_1^3.

We measure the applicability of a family of space-filling curves based upon their common structural characteristics, which are informally described as follows. Locality preservation reflects proximity between the grid points of $[n]^m$, that is, close-by points in $[n]^m$ are mapped to close-by indices/numbers in $[n^m]$, or vice versa. Clustering performance measures the distribution of continuous runs of grid points (clusters) over all identically shaped subspaces of $[n]^m$, which can be characterized by the mean number of clusters and the mean inter-cluster distance (in $[n^m]$) within a subspace.

A few locality measures have been proposed and analyzed for space-filling curves in the literature (see [MD86], [GL96], [NRS97], [Alb97], [AN00], and [DS03]). Different measures are defined to address the proximity preservation of close-by points in the m-dimensional grid space $[n]^m$ or in the indexing space $[n^m]$. Generally, Hilbert curve family, z-order curve family, and H-indexings [NRS97] achieve good locality performances.

Empirical and analytical studies of clustering performances of various low-dimensional space-filling curves have been reported in the literature (see [Jag97] and [MJFS01] for details). Generally, the Hilbert curve family exhibits good performance in these studies.

Jagadish [Jag97] derives exact formulas for the mean numbers of clusters over all rectangular 2×2 and 3×3 subgrids of an H_k^2-structural grid space. Moon, Jagadish, Faloutsos, and Saltz [MJFS01] prove that in a sufficiently large m-dimensional H_k^m-structural grid space, the mean number of clusters over all rectilinear polyhedral queries with surface area $S_{m,k}$ approaches $\frac{1}{2}\frac{S_{m,k}}{m}$ as k approaches ∞. They also extend the work in [Jag97] to obtain the exact formula for the mean number of clusters over all rectangular $2^q \times 2^q$ subgrids of an H_k^2-structural grid space.

This paper presents an application of random walk to the study of inter-clustering of space-filling curves and an analytical study on the inter-clustering performances of 2-dimensional Hilbert and z-order curve families. For an m-dimensional space-filling curve $C : [n^m] \to [n]^m$ and a subgrid G of $[n]^m$, a cluster of G induced by C is a maximal (contiguous) subinterval I of $[n^m]$ such that $C(I) \subseteq G$. We can partition and order $C^{-1}(G)$ into disjoint union of clusters. An inter-cluster gap of G is a subinterval of $[n^m]$ delimited by two consecutive clusters of G, and the corresponding inter-cluster distance is the length of the inter-cluster gap. Thus, the space-filling curve C induces the following statistics: (1) the mean number of clusters of

$C^{-1}(G)$ over all identically shaped subgrids G of $[n]^m$, (2) the (universe) mean inter-cluster distance over all inter-cluster gaps from all identically shaped subgrids G of $[n]^m$, and (3) the mean total inter-cluster distance (in a subgrid) over all identically shaped subgrids G of $[n]^m$.

The studies of clustering and inter-clustering performances for space-filling curves are motivated by the applicability of multi-dimensional space-filling indexing methods, in which an m-dimensional data space is mapped onto a 1-dimensional data space (external storage structure) by adopting a 1-dimensional indexing method based upon an m-dimensional space-filling curve.

The space-filling index structure can support efficient query processing (such as range queries) provided that we minimize the average number of external fetch/seek operations, which is related to the clustering statistics. Asano, Ranjan, Roos, Welzl, and Widmayer [ARR$^+$97] study the optimization of range queries over space-filling index structures, which aims at minimizing the number of seek operations (not the number of block accesses) — trade-off between seek time to proper block (cluster) and latency/transfer time for unnecessary blocks (inter-cluster gap). Good bounds on the two inter-clustering statistics translate into good bounds on the average tolerance of unnecessary block transfers.

We show how approximating the mean inter-cluster distance statistics of continuous multi-dimensional space-filling curves fits into the formalism of random walk, and derive exact formulas for the two inter-clustering statistics for 2-dimensional Hilbert and z-order curve families over all identically shaped square subgrids of $[n]^2$, with computer program verification over various grid- and subgrid-orders. Our comparisons are accordingly twofold: first to gauge the relative performances of the two curve families with respect to the two inter-clustering statistics based upon the analytical results, and second, to check the applicability of the random-walk approximation to the universe mean inter-clustering distance based upon the approximation and analytical results. Note that we present the skeletons for proving the main results without the lengthly derivations. Complete proofs and verifying programs are available from the authors.

2 Approximation with Random Walks

Consider an m-dimensional continuous space-filling curve $C : [n^m] \to [n]^m$. Denote the frequency distribution of edge-direction of C with respect to the m-dimensional Cartesian coordinates by $(d_i)_{i=1}^m$. Note that in a typical application of an m-dimensional order-k space-filling curve of length n^m ($n = 2^k$), k (hence n) is sufficiently large. We derive our statistical/approximation application of an m-dimensional random walk in the absence of grid-boundaries.

The principal random elements defining the random walk are the successive unit-step transitions (since C is continuous) from a grid point to one of its $2m$ neighboring grid points according to the edge-direction distribution: $(p_{i^+}, p_{i^-})_{i=1}^m$, where p_{i^+} and p_{i^-} denote the transition probabilities in the positive (i^+) and negative (i^-) ith-axis directions, respectively, with $p_{i^+} + p_{i^-} = d_i$ for $i = 1, 2, \ldots, m$. Denote by $p_{i^\perp}$ the probability of a one-step transition orthogonal to the ith-axis; thus $p_{i^\perp} = \sum_{j \mid j \neq i}(p_{j^+} + p_{j^-}) = 1 - d_i$.

Let G be a hyperrectangular query subgrid of $[n]^m$, and we consider how an inter-cluster gap J evolves in our random-walk context, starting at a grid point (in $[n]^m - G$) neighboring a boundary hyperplane P of G. Assume for computational simplicity that J (first) returns into G through P. Without loss of generality, assume that the normal of P is the jth-axis, and the first return of J into G through P is in the j^--direction.

Consider the event "$|J| = \gamma$" — length of J is γ for some positive integer γ and its probability. The ordered sequence of transitions of J embeds a subsequence J' such that:

1. J' consists of all j^+- and j^--transitions of J and terminated with the first-return j^--transition of J into G through P. Equivalently, $J - J'$ consists of all $j^\perp$-transitions (parallel to P) of J, and

2. The subsequence J'' of J' excluding the first-return j^--transition of J exhibits the Catalan structure (see [GKP94]):
   (a) number of j^+-transitions of J'' = number of j^--transitions of J'', and
   (b) For every proper prefix J''' of J'', number of j^+-transitions of $J''' >$ number of j^--transitions of J'''.

Thus, we have, for all positive integers γ,

$$\Pr(|J| = \gamma + 1) = (\sum_{l \geq 0} \binom{\gamma}{2l} c_l\, p_{j^+}^l\, p_{j^-}^l\, p_{j^\perp}^{\gamma - 2l})\, p_{j^-},$$

where c_l denotes the Catalan number $\binom{2l}{l}\frac{1}{l+1}$.

For computational simplicity, assume that the underlying random walk is symmetric with respect to each ith-axis for $i = 1, 2, \ldots, m$; that is, $p_{i^+} = p_{i^-} = \frac{d_i}{2}$. The probability above becomes:

$$\sum_{l \geq 0} \binom{\gamma}{2l} c_l (\frac{d_j}{2})^{2l+1} (1 - d_j)^{\gamma - 2l}.$$

An m-dimensional Hilbert curve enjoys a uniformly distributed $(d_i)_{i=1}^m$ asymptotically, and we can express the probability above and an approximate mean inter-cluster distance statistics in terms of some well-known functions.

Lemma 1 *For an m-dimensional Hilbert curve of length n^m with its edge-direction distribution $(d_i)_{i=1}^m$, $\lim_{n\to\infty} \frac{d_i}{d_{i'}} = 1$ for all $i, i' \in \{1, 2, \ldots, m\}$.*

Let F denote the hypergeometric function (see [GKP94]): $F\left(\begin{matrix} a_1, \ldots, a_p \\ b_1, \ldots, b_q \end{matrix} \;\middle|\; z\right) = \sum_{k \geq 0} \frac{a_1^{\overline{k}} \cdots a_p^{\overline{k}}}{b_1^{\overline{k}} \cdots b_q^{\overline{k}}} \frac{z^k}{k!}$ with upper parameters a's and lower parameters b's, where $x^{\overline{k}}$ denotes the rising factorial power.

Lemma 2 *For all positive integers $\gamma \geq 2$,*

$$Pr(|J| = \gamma) = \sum_{l \geq 0} \binom{\gamma - 1}{2l} c_l (\frac{1}{2m})^{2l+1} (\frac{m-1}{m})^{\gamma - 1 - 2l} = F\left(\begin{matrix} -\frac{\gamma}{2} + \frac{1}{2}, -\frac{\gamma}{2} + 1 \\ 2 \end{matrix} \;\middle|\; \frac{1}{(m-1)^2}\right) \frac{1}{2m} (\frac{m-1}{m})^{\gamma - 1}.$$

For an m-dimensional Hilbert curve of length $N = n^m$, the random walk formulated above yields an approximate mean inter-cluster distance statistics based upon $\sum_{\gamma=1}^N \gamma \Pr(|J| = \gamma)$. To measure the goodness of our approximation model versus an analytical study presented below, we consider the case of $m = 2$ (see [BRWW97]). Let Γ denote the Gamma function.

Lemma 3 *For a 2-dimensional Hilbert curve of length $N = n^2$,*

1. *The inter-cluster distance probability is:*

$$Pr(|J| = \gamma) = \frac{\Gamma(\gamma + \frac{1}{2})}{\sqrt{\pi}\,\Gamma(\gamma + 2)} = \frac{c_\gamma}{4^\gamma}.$$

2. *The approximate mean inter-cluster distance statistics is:*

$$\sum_{\gamma=1}^{N} \gamma Pr(|J| = \gamma) = \frac{2(N+2)^2\Gamma(N+\frac{3}{2})}{\sqrt{\pi}\Gamma(N+3)} - 2 = \frac{(N+2)(2N+1)}{4^N} c_N - 2.$$

Note that $c_k \sim \frac{4^k}{\sqrt{\pi}k^{\frac{3}{2}}}$ by using Stirling's formula. Thus the approximate mean inter-cluster distance for 2-dimensional Hilbert curve of length $N = n^2$ is asymptotically $\frac{2}{\sqrt{\pi}} N^{\frac{1}{2}}$ $(= \frac{2}{\sqrt{\pi}} n)$.

3 Analytical Study of Inter-clustering Performances

Our analytical study of inter-clustering performances is focused on 2-dimensional Hilbert and z-order curve families. We develop and state all supporting lemmas for the Hilbert curve family in this section; those for the z-order curve family can be obtained analogously.

For a mathematical formalism of discrete Hilbert curves that facilitates combinatorial studies of multi-dimensional Hilbert indexing, see [AN00] for details. One of the salient characteristics of Hilbert curves is their "self-similarity" — a Hilbert curve can be generated by interconnecting identical subcurves via reflection and rotation (see Figure 2). For 2-dimensional Hilbert curves, this self-similar structural property guides us to decompose H_k^2 into four identical H_{k-1}^2-subcurves (via reflection and rotation), which are amalgamated together by an H_1^2-curve. Following the linear order along this H_1^2-curve, we denote the four H_{k-1}^2-subcurves as $Q_1(H_k^2)$, $Q_2(H_k^2)$, $Q_3(H_k^2)$, and $Q_4(H_k^2)$.

For a 2-dimensional grid, the "orientation" of H_k^2 uniquely determines that of $Q_\alpha(H_k^2)$ for $\alpha = 1, 2, 3, 4$, and thus only one H_k^2 exists modulo symmetry (whereas there are 1536 structurally different 3-dimensional Hilbert curves [AN00]). For a 2-dimensional Hilbert curve H_k^2 indexing the grid $[2^k]^2$, with a canonical orientation shown in Figure 2(a), we denote by $\partial_1(H_k^2)$ and $\partial_2(H_k^2)$ the entry and exit, respectively, grid point in $[2^k]^2$ (with respect to the canonical orientation). Figure 2 depicts the decomposition of H_k^2 and the ∂_1- and ∂_2-labels of four H_{k-1}^2-subcurves.

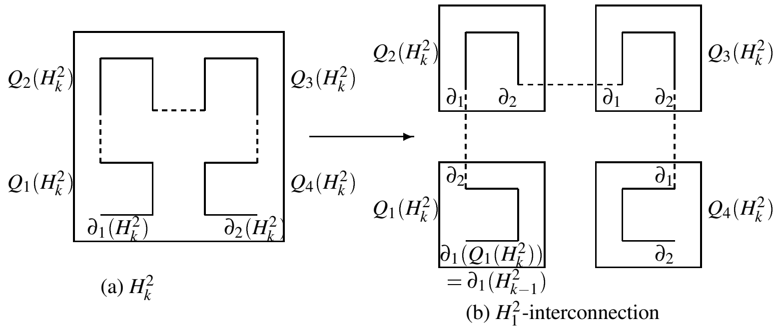

Fig. 2: Generation of H_k^2 in (a) from a H_1^2-interconnection of four H_{k-1}^2-subcurves in (b).

With respect to the canonical orientation of H_k^2 shown in Figure 2(a), we cover the 2-dimensional k-order grid with 2^k rows $(R_{k,1}, R_{k,2}, \ldots, R_{k,2^k})$, indexed from the bottom, and 2^k columns $(C_{k,1}, C_{k,2}, \ldots, C_{k,2^k})$, indexed from the left. We denote:

1. For a grid point $v \in [2^k]^2$, its x- and y-coordinate by $X(v)$ and $Y(v)$, respectively (that is, v is the intersection grid point of the column $C_{k,X(v)}$ and the row $R_{k,Y(v)}$),

2. For the grid points $v, v' \in [2^k]^2$, their index-difference by $\hbar_k(v,v')$ $(= |(H_k^2)^{-1}(v) - (H_k^2)^{-1}(v')|)$, and

3. For a rectangular query subgrid with its lower-left corner at grid point (x,y) and upper-right corner at grid point (x',y') $(1 \le x \le x' \le 2^k$ and $1 \le y \le y' \le 2^k)$ covering $\cup_{\alpha=x}^{x'} C_{k,\alpha} \cap \cup_{\beta=y}^{y'} R_{k,\beta}$, its set of grid points by $G_k(x,y,x',y')$ $(= \{v \in [2^k]^2 \mid x \le X(v) \le x'$ and $y \le Y(v) \le y'\})$. The size of the query subgrid $G_k(x,y,x',y')$ is $(x'-x+1) \times (y'-y+1)$.

Remark 1. *For most self-similar m-dimensional order-k space-filling curve C_k^m indexing the grid $[2^k]^m$, we can view C_k^m as a C_{k-q}^m-curve interconnecting $2^{2(k-q)}$ C_q^m-subcurves for all $q \in [k]$.*

The remark above motivates our analytical study of inter-clustering performances to be based upon query subgrids of size $2^q \times 2^q$.

For a 2-dimensional order-k Hilbert curve H_k^2, let $\Psi_q(H_k^2)$ denote the summation of all inter-cluster distances over all $2^q \times 2^q$ query subgrids of an H_k^2-structural grid space $[2^k]^2$. For a subgrid G, let $\theta_1(G)$ denote the first entrance (the lowest H_k^2-indexed grid point) into G and $\theta_2(G)$ denote the last exit (the highest H_k^2-indexed grid point) out of G.

Remark 2. *Within a query subgrid G (with $|G|$ grid points), the summation of all its inter-cluster distances is $\hbar_k(\theta_1(G),\theta_2(G)) - |G| + 1$. In developing the supporting lemmas, we express $\hbar_k(\theta_1(G),\theta_2(G))$ as $\hbar_k(\theta_2(G),v) - \hbar_k(\theta_1(G),v)$ for a suitably chosen grid point v.*

Remark 2 reduces the computation of the summation of all inter-cluster distances over all identically shaped subgrids G to the computations of $\sum_{\text{all } G} \hbar_k(\theta_j(G),v)$ for $j = 1,2$ and a suitably chosen v.

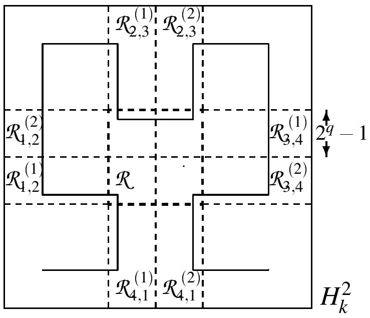

Fig. 3: The boundary regions of neighboring quadrants are organized into nine disjoint regions: $\mathcal{R}_{i,i \bmod 4+1}^{(1)}$, $\mathcal{R}_{i,i \bmod 4+1}^{(2)}$ for $i = 1,2,3,4$, and $\mathcal{R}$.

The recursive decomposition of H_k^2 (see Figure 2(b)) gives that

$$\Psi_q(H_k^2) = 4\Psi_q(H_{k-1}^2) + \varepsilon_{k,q}(H_k^2),$$

where $\varepsilon_{k,q}(H_k^2)$ denotes the summation of all inter-cluster distances over all $2^q \times 2^q$ query subgrids, each of which overlaps with more than one quadrant (that is, two or four). These query subgrids are contained in the boundary regions of neighboring quadrants, which can be organized into nine disjoint regions: $\mathcal{R}_{i,i \bmod 4+1}^{(1)}$, $\mathcal{R}_{i,i \bmod 4+1}^{(2)}$ for $i = 1,2,3,4$, and $\mathcal{R}$, as shown in Figure 3.

Remark 3. *For a query subgrid G overlapping with more than one quadrant, $\theta_1(G)$ is in the lowest-numbered quadrant, and $\theta_2(G)$ is in the highest-numbered quadrant.*

For a $2^q \times 2^q$ query subgrid G, G overlaps with:

1. Exactly $Q_i(H_k^2)$ and $Q_{i \bmod 4+1}(H_k^2)$ if and only if $G \subseteq \mathcal{R}_{i,i \bmod 4+1}^{(1)} \cup \mathcal{R}_{i,i \bmod 4+1}^{(2)}$ for every $i \in \{1,2,3,4\}$. In this case, $\theta_j(G) \in \mathcal{R}_{i,i \bmod 4+1}^{(j)}$ for $j \in \{1,2\}$ by Remark 3.

2. $Q_i(H_k^2)$ for all $i \in \{1,2,3,4\}$ if and only if $G \subseteq \mathcal{R}$. In this case, $\theta_1(G) \in Q_1(H_k^2)$ (upper-right corner) and $\theta_2(G) \in Q_4(H_k^2)$ (upper-left corner) by Remark 3.

We divide the computation of $\varepsilon_{k,q}(H_k^2)$ into three parts:

1. $\sum \hbar_k(\theta_2(G), \partial_1(H_k^2))$ over all $2^q \times 2^q$ query subgrids $G \subseteq \mathcal{R}_{i,i \bmod 4+1}^{(1)} \cup \mathcal{R}_{i,i \bmod 4+1}^{(2)}$ for $i \in \{1,2,3,4\}$,

2. $\sum \hbar_k(\theta_1(G), \partial_1(H_k^2))$ over all $2^q \times 2^q$ query subgrids $G \subseteq \mathcal{R}_{i,i \bmod 4+1}^{(1)} \cup \mathcal{R}_{i,i \bmod 4+1}^{(2)}$ for $i \in \{1,2,3,4\}$, and

3. the summation of all inter-cluster distances over all $2^q \times 2^q$ query subgrids contained in $\mathcal{R}$.

We develop combinatorial lemmas in the following three subsections to support the computations.

3.1 $\sum \hbar_k(\theta_2(G), \partial_1(H_k^2))$ over Subgrids G Overlapping with Two Quadrants

Consider an arbitrary $2^q \times 2^q$ query subgrid $G \subseteq \mathcal{R}_{i,i \bmod 4+1}^{(1)} \cup \mathcal{R}_{i,i \bmod 4+1}^{(2)}$ where $i \in \{1,2,3,4\}$. Remark 3 gives that $\theta_2(G) \in \mathcal{R}_{i,i \bmod 4+1}^{(2)}$, and we zoom in on the "incomplete" rectangular subgrid $G \cap \mathcal{R}_{i,i \bmod 4+1}^{(2)}$ (with one side-length at most $2^q - 1$). Observe that for $i = 1,2,3,4$, $\mathcal{R}_{i,i \bmod 4+1}^{(2)}$ aggregates the $2^q - 1$ bottom rows, leftmost columns, top rows, and leftmost columns of $Q_2(H_k^2)$, $Q_3(H_k^2)$, $Q_4(H_k^2)$, and $Q_4(H_k^2)$, respectively. Since the quadrants are isomorphic to a canonical H_{k-1}^2 via symmetry (reflection and rotation), we consider the following system of summations $\Omega_{k,2^q} = (\Omega_{k,2^q}^L, \Omega_{k,2^q}^R, \Omega_{k,2^q}^B, \Omega_{k,2^q}^T)$ in a general context of a canonical H_k^2:

$$\begin{aligned}
\Omega^L_{k,2^q} &= \sum_{x=1}^{2^q-1}\sum_{y=1}^{2^k-2^q+1} \hbar_k(\theta_2(G_k(1,y,x,y+2^q-1)),\partial_1(H_k^2)) \text{ — for left boundary (see Figure 4(a)),}\\
\Omega^R_{k,2^q} &= \sum_{x=2^k-2^q+2}^{2^k}\sum_{y=1}^{2^k-2^q+1} \hbar_k(\theta_2(G_k(x,y,2^k,y+2^q-1)),\partial_1(H_k^2)) \text{ — for right boundary,}\\
\Omega^B_{k,2^q} &= \sum_{x=1}^{2^k-2^q+1}\sum_{y=1}^{2^q-1} \hbar_k(\theta_2(G_k(x,1,x+2^q-1,y)),\partial_1(H_k^2)) \text{ — for bottom boundary,}\\
\Omega^T_{k,2^q} &= \sum_{x=1}^{2^k-2^q+1}\sum_{y=2^k-2^q+2}^{2^k} \hbar_k(\theta_2(G_k(x,y,x+2^q-1,2^k)),\partial_1(H_k^2)) \text{ — for top boundary, and}\\
\mathcal{N}^S_{k,2^q} &= \sum_{x=1}^{2^q-1}\sum_{y=1}^{2^k-2^q+1} 1 \text{ — for the number of incomplete rectangular subgrids in a boundary.}
\end{aligned}$$

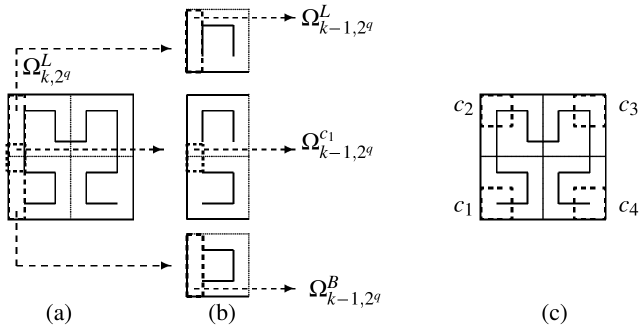

Fig. 4: (a) $\Omega^L_{k,2^q}$ for a canonical H_k^2; (b) its recursive decomposition; (c) the four $(2^q-1)\times(2^q-1)$ corners of a canonical H_k^2.

We will establish a system of recurrences (in k) for $\Omega_{k,2^q}$ (see Lemma 7 below). The system of recurrence involves another system of summations as prerequisites, as demonstrated in the following example. Consider a recursive decomposition of $\Omega^L_{k,2^q}$, illustrated in Figure 4(a) and (b), into four parts: (1) $\Omega^B_{k-1,2^q}$, (2) $\Omega^{c_1}_{k-1,2^q}$, (3) $\Omega^L_{k-1,2^q}$, and (4) adjustments for the previous three parts. The part $\Omega^{c_1}_{k-1,2^q}$ helps compute $\sum \hbar_k(\theta_2(G),\partial_1(H_k^2))$ over all incomplete rectangular subgrids G (with one side-length at most 2^q-1) overlapping both $Q_1(H_k^2)$ and $Q_2(H_k^2)$. According to Remark 3, the computation of this summation is reduced to $\sum \hbar_{k-1}(\theta_2(G),\partial_1(H_{k-1}^2))$ over all incomplete rectangular subgrids G (with both side-lengths at most 2^q-1) in the c_1-corner (lower-left corner) of a canonical H_{k-1}^2 (that is, $Q_2(H_k^2)$). Each of the three parts $\Omega^B_{k-1,2^q}$, $\Omega^{c_1}_{k-1,2^q}$, and $\Omega^L_{k-1,2^q}$ is defined with respect to $\partial_1(H_{k-1}^2)$ of a canonical H_{k-1}^2, we need to adjust each part with distance cumulation between the entry/exit of the underlying quadrant and $\partial_1(H_k^2)$.

The recursive decompositions of all four parts in $\Omega^L_{k,2^q}$, $\Omega^R_{k,2^q}$, $\Omega^B_{k,2^q}$, and $\Omega^T_{k,2^q}$ lead us to consider a prerequisite system of summations $\Omega^c_{k,2^q} = (\Omega^{c_1}_{k,2^q},\Omega^{c_2}_{k,2^q},\Omega^{c_3}_{k,2^q},\Omega^{c_4}_{k,2^q})$ in a more general context of a canonical H_k^2 (see Figure 4(c)):

$$\begin{aligned}
\Omega_{k,2^q}^{c_1} &= \sum_{x=1}^{2^q-1}\sum_{y=1}^{2^q-1} \hbar_k(\theta_2(G_k(1,1,x,y)),\partial_1(H_k^2)) \text{ — for lower-left corner,}\\
\Omega_{k,2^q}^{c_2} &= \sum_{x=1}^{2^q-1}\sum_{y=2^k-2^q+2}^{2^k} \hbar_k(\theta_2(G_k(1,y,x,2^k)),\partial_1(H_k^2)) \text{ — for upper-left corner,}\\
\Omega_{k,2^q}^{c_3} &= \sum_{x=2^k-2^q+2}^{2^k}\sum_{y=2^k-2^q+2}^{2^k} \hbar_k(\theta_2(G_k(x,y,2^k,2^k)),\partial_1(H_k^2)) \text{ — for upper-right corner,}\\
\Omega_{k,2^q}^{c_4} &= \sum_{x=2^k-2^q+2}^{2^k}\sum_{y=1}^{2^q-1} \hbar_k(\theta_2(G_k(x,1,2^k,y)),\partial_1(H_k^2)) \text{ — for lower-right corner, and}\\
\mathcal{N}_{k,2^q}^{c} &= \sum_{x=1}^{2^q-1}\sum_{y=1}^{2^q-1} 1 \text{ — for the number of incomplete rectangular subgrids in a corner.}
\end{aligned}$$

Note that in $\Omega_{k,2^q}^{c_4}$, $\theta_2(G_k(x,1,2^k,y)) = \partial_2(H_k^2)$ for all x and y in the summation-index ranges, hence $\Omega_{k,2^q}^{c_4} = (2^q-1)^2\hbar_k(\partial_2(H_k^2),\partial_1(H_k^2)) = (2^q-1)^2(2^{2k}-1)$. All other three summations involve rectangular subgrids contained in $(2^q-1)\times(2^q-1)$ corners. As suggested by Remark 1, we zoom in on the $2^q\times 2^q$ H_q^2-structural corners, and consider the following system of summations $\overline{\Omega}_{q,2^q}^{c} = (\overline{\Omega}_{q,2^q}^{c_1}, \overline{\Omega}_{q,2^q}^{c_2}, \overline{\Omega}_{q,2^q}^{c_3})$:

$$\begin{aligned}
\overline{\Omega}_{q,2^q}^{c_1} &= \sum_{x=1}^{2^q}\sum_{y=1}^{2^q} \hbar_q(\theta_2(G_q(1,1,x,y)),\partial_1(H_q^2)) \text{ — for lower-left corner,}\\
\overline{\Omega}_{q,2^q}^{c_2} &= \sum_{x=1}^{2^q}\sum_{y=1}^{2^q} \hbar_q(\theta_2(G_q(1,y,x,2^q)),\partial_1(H_q^2)) \text{ — for upper-left corner,}\\
\overline{\Omega}_{q,2^q}^{c_3} &= \sum_{x=1}^{2^q}\sum_{y=1}^{2^q} \hbar_q(\theta_2(G_q(x,y,2^q,2^q)),\partial_1(H_q^2)) \text{ — for upper-right corner, and}\\
\overline{\mathcal{N}}_{q,2^q}^{c} &= \sum_{x=1}^{2^q}\sum_{y=1}^{2^q} 1 \text{ — for the number of rectangular subgrids in a } 2^q\times 2^q \text{ corner.}
\end{aligned}$$

Thus far, we learn that the system of recurrences for $\Omega_{k,2^q}$ can be defined and solved via the prerequisite system $\Omega_{k,2^q}^{c}$, which is related to the system $\overline{\Omega}_{q,2^q}^{c}$ (see Lemma 6 below). The system $\overline{\Omega}_{q,2^q}^{c}$, which involves subgrids (with both side-lengths at most 2^q) of a canonical H_q^2, represents the basis of the recursive decompositions (in k to q) of $\Omega_{k,2^q}$ and $\Omega_{k,2^q}^{c}$. Similar to the reduction of $\Omega_{k,2^q}$ to $\Omega_{k,2^q}^{c}$, we develop a system of recurrences (in q) for $\overline{\Omega}_{q,2^q}^{c}$ via a prerequisite system, as demonstrated in the following example. Consider a recursive decomposition of $\overline{\Omega}_{q,2^q}^{c_1} = \sum_{x=1}^{2^q}\sum_{y=1}^{2^q}\hbar_k(\theta_2(G_q(1,1,x,y)),\partial_1(H_q^2))$ into four parts (together with adjustments), based upon the overlapping scenario of the rectangular subgrid $G_q(1,1,x,y)$ with the four quadrants of a canonical H_q^2 (see Figure 5).

Case 1: $G_q(1,1,x,y)$ is contained in $Q_1(H_q^2)$ (see Figure 5(a)). This part is reduced to $\overline{\Omega}_{q-1,2^{q-1}}^{c_1}$ after $(+\frac{\pi}{2})$-rotating and then reflecting $Q_1(H_q^2)$ into a canonical H_{q-1}^2.

Case 2: $G_q(1,1,x,y)$ overlaps with exactly $Q_1(H_q^2)$ and $Q_2(H_q^2)$ (see Figure 5(b)). This part is reduced to $\overline{\Omega}_{q-1,2^{q-1}}^{c_1}$ (with adjustment of distance cumulation).

Case 3: $G_q(1,1,x,y)$ overlaps with exactly $Q_1(H_q^2)$ and $Q_4(H_q^2)$ (see Figure 5(d)). This part is reduced to $\overline{\Omega}^{c_3}_{q-1,2^{q-1}}$ after $(-\frac{\pi}{2})$-rotating and then reflecting $Q_4(H_q^2)$ into a canonical H_{q-1}^2 (with adjustment of distance cumulation).

Case 4: $G_q(1,1,x,y)$ overlaps with all quadrants (see Figure 5(c)). The overlapping condition gives that $x,y \in \{2^{q-1}+1, 2^{q-1}+2, \ldots, 2^q\}$. According to Remark 3, $\theta_2(G_q(1,1,x,y)) \in Q_4(H_q^2)$. Observe that, as shown in Figure 6, for every $x \in \{2^{q-1}+1, 2^{q-1}+2, \ldots, 2^q\}$, the subgrids $G_q(1,1,x,y)$ for all $y \in \{2^{q-1}+1, 2^{q-1}+2, \ldots, 2^q\}$ have the same $\theta_2(G_q(1,1,x,y))$ (independent of y).

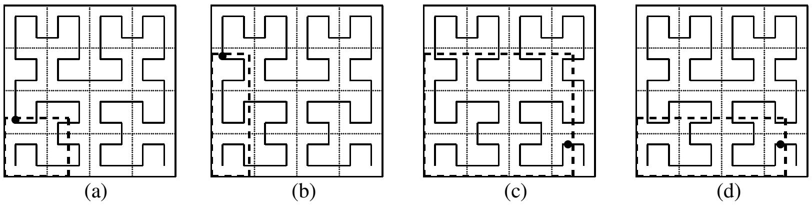

Fig. 5: Four overlapping scenarios when decomposing $\overline{\Omega}^{c_1}_{q,2^q}$ in a canonical H_q^2: (a) contained in $Q_1(H_q^2)$; (b) and (d) overlapping with exactly two quadrants; (c) overlapping with all quadrants.

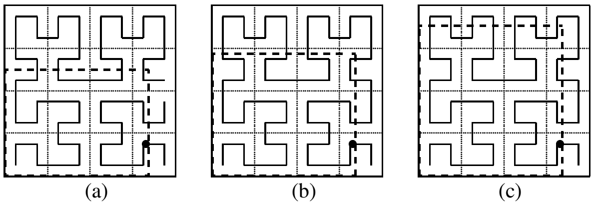

Fig. 6: For subgrids overlapping with all quadrants of a canonical H_q^2, their last exits are the same.

The recursive decompositions of $\overline{\Omega}^{c_1}_{q,2^q}$, $\overline{\Omega}^{c_2}_{q,2^q}$, and $\overline{\Omega}^{c_3}_{q,2^q}$ lead us to consider a prerequisite system of summations $\overline{\Pi}_q = (\overline{\Pi}^T_q, \overline{\Pi}^L_q)$ in a general context of a canonical H_q^2:

$$\begin{aligned}
\overline{\Pi}^T_q &= \sum_{y=2^q}^{1} \hbar_q(\theta_2(G_q(1,y,2^q,2^q)), \partial_1(H_q^2)) \text{ — top to bottom incrementally, and} \\
\overline{\Pi}^L_q &= \sum_{x=1}^{2^q} \hbar_q(\theta_2(G_q(1,1,x,2^q)), \partial_1(H_q^2)) \text{ — left to right incrementally.}
\end{aligned}$$

We develop and solve a system of recurrences for $\overline{\Pi}_q$ and reverse the sequence of reductions to obtain the closed-form solutions for $\Omega_{k,2^q}$, which are summarized in the following four lemmas. Note that we present the systems of recurrences only (which are solved by a mathematical and analytical software such as Maple).

Lemma 4 *For a canonical H_q^2,*

$$\overline{\Pi}_q^T = \begin{cases} \overline{\Pi}_{q-1}^T + 2(2^{q-1})^3 + \overline{\Pi}_{q-1}^L + 3(2^{q-1})^3 & \text{if } q > 1 \\ 5 & \text{if } q = 1 \end{cases}$$

$$\overline{\Pi}_q^L = \begin{cases} \overline{\Pi}_{q-1}^L + (2^{q-1})^3 + \overline{\Pi}_{q-1}^T + 3(2^{q-1})^3 & \text{if } q > 1 \\ 4 & \text{if } q = 1 \end{cases}$$

The closed-form solutions for $\overline{\Pi}_q$ are employed to establish a system of recurrences for $\overline{\Omega}_{q,2^q}^c$.

Lemma 5 *For a canonical H_q^2,*

$$\overline{\Omega}_{q,2^q}^{c_1} = \begin{cases} 2\overline{\Omega}_{q-1,2^{q-1}}^{c_1} + \overline{\Omega}_{q-1,2^{q-1}}^{c_3} + \frac{5^3}{2^8} \cdot 2^{4q} - \frac{3}{2^4} \cdot 2^{2q} & \text{if } q > 1, \\ 7 & \text{if } q = 1; \end{cases}$$

$$\overline{\Omega}_{q,2^q}^{c_2} = \begin{cases} 3\overline{\Omega}_{q-1,2^{q-1}}^{c_2} + \frac{3 \cdot 41}{2^8} \cdot 2^{4q} - \frac{3}{2^4} \cdot 2^{2q} & \text{if } q > 1, \\ 7 & \text{if } q = 1; \end{cases}$$

$$\overline{\Omega}_{q,2^q}^{c_3} = \begin{cases} \overline{\Omega}_{q-1,2^{q-1}}^{c_1} + \overline{\Omega}_{q-1,2^{q-1}}^{c_3} + \frac{23}{2^5} \cdot 2^{4q} - \frac{3}{2^3} \cdot 2^{2q} & \text{if } q > 1, \\ 10 & \text{if } q > 1. \end{cases}$$

The closed-form solutions for $\overline{\Omega}_{q,2^q}^c$ and $\overline{\Pi}_q$ are employed to obtain exact formulas for $\Omega_{k,2^q}^c$.

Lemma 6 *For a canonical H_k^2 structured as an H_{k-q}^2-curve interconnecting $2^{2(k-q)}$ H_q^2-subcurves,*

$$\begin{aligned} \Omega_{k,2^q}^{c_1} &= \overline{\Omega}_{q,2^q}^{c_1} - \overline{\Pi}_q^L - (2^q - 1)(2^{2q} - 1), \\ \Omega_{k,2^q}^{c_2} &= \overline{\Omega}_{q,2^q}^{c_2} - \overline{\Pi}_q^T - \overline{\Pi}_q^L + (2^{2q} - 1) + (2^q - 1)^2 \sum_{i=q}^{k-1} 2^{2i}, \\ \Omega_{k,2^q}^{c_3} &= \overline{\Omega}_{q,2^q}^{c_3} - \overline{\Pi}_q^T - (2^q - 1)(2^{2q} - 1) + (2^q - 1)^2 (2^{2k} - \sum_{i=q}^{k-1} 2^{2i} - 2^{2q}). \end{aligned}$$

The exact formulas for $\Omega_{k,2^q}^c$ are employed to establish a system of recurrences for $\Omega_{k,2^q}$.

Lemma 7 *For a canonical H_k^2 structured as an H_{k-q}^2-curve interconnecting $2^{2(k-q)}$ H_q^2-subcurves,*

$$\Omega_{k,2^q}^L = \begin{cases} \Omega_{k-1,2^q}^B + (\Omega_{k-1,2^q}^{c_1} + (2^q - 1)^2 (2^{k-1})^2) + (\Omega_{k-1,2^q}^L + (2^{k-1} - 2^q + 1)(2^q - 1)(2^{k-1})^2) & \text{if } k > q, \\ \overline{\Pi}_q^L - (2^{2q} - 1) & \text{if } k = q; \end{cases}$$

$$\Omega_{k,2^q}^R = \begin{cases} (\Omega_{k-1,2^q}^B + 3(2^{k-1} - 2^q + 1)(2^q - 1)(2^{k-1})^2) + (\Omega_{k-1,2^q}^{c_1} + 3(2^q - 1)^2 (2^{k-1})^2) \\ \quad + (\Omega_{k-1,2^q}^R + 2(2^{k-1} - 2^q + 1)(2^q - 1)(2^{k-1})^2) & \text{if } k > q, \\ (2^q - 1)(2^{2q} - 1) & \text{if } k = q; \end{cases}$$

$$\Omega_{k,2^q}^B = \begin{cases} \Omega_{k-1,2^q}^L + (\Omega_{k-1,2^q}^{c_3} + 3(2^q - 1)^2 (2^{k-1})^2) + (\Omega_{k-1,2^q}^R + 3(2^{k-1} - 2^q + 1)(2^q - 1)(2^{k-1})^2) & \text{if } k > q, \\ (2^q - 1)(2^{2q} - 1) & \text{if } k = q; \end{cases}$$

$$\Omega_{k,2^q}^T = \begin{cases} (\Omega_{k-1,2^q}^T + (2^{k-1} - 2^q + 1)(2^q - 1)(2^{k-1})^2) + (\Omega_{k-1,2^q}^{c_2} + 2(2^q - 1)^2 (2^{k-1})^2) \\ \quad + (\Omega_{k-1,2^q}^T + 2(2^{k-1} - 2^q + 1)(2^q - 1)(2^{k-1})^2) & \text{if } k > q, \\ \overline{\Pi}_q^T - (2^{2q} - 1) & \text{if } k = q. \end{cases}$$

We obtain the closed-form solutions for $\Omega_{k,2^q}$ by using the mathematical software Maple.

3.2 $\sum \hbar_k(\theta_1(G), \partial_1(H_k^2))$ over Subgrids G Overlapping with Two Quadrants

We may proceed as in Section 3.1, based upon the system of summations $\omega_{k,2^q} = (\omega_{k,2^q}^L, \omega_{k,2^q}^R, \omega_{k,2^q}^B, \omega_{k,2^q}^T)$:

$$\begin{aligned}
\omega_{k,2^q}^L &= \sum_{x=1}^{2^q-1} \sum_{y=1}^{2^k-2^q+1} \hbar_k(\theta_1(G_k(1,y,x,y+2^q-1)), \partial_1(H_k^2)) \text{ — for left boundary,} \\
\omega_{k,2^q}^R &= \sum_{x=2^k-2^q+2}^{2^k} \sum_{y=1}^{2^k-2^q+1} \hbar_k(\theta_1(G_k(x,y,2^k,y+2^q-1)), \partial_1(H_k^2)) \text{ — for right boundary,} \\
\omega_{k,2^q}^B &= \sum_{x=1}^{2^k-2^q+1} \sum_{y=1}^{2^q-1} \hbar_k(\theta_1(G_k(x,1,x+2^q-1,y)), \partial_1(H_k^2)) \text{ — for bottom boundary, and} \\
\omega_{k,2^q}^T &= \sum_{x=1}^{2^k-2^q+1} \sum_{y=2^k-2^q+2}^{2^k} \hbar_k(\theta_1(G_k(x,y,x+2^q-1,2^k)), \partial_1(H_k^2)) \text{ — for top boundary.}
\end{aligned}$$

Or, we apply the following lemma to relate the two systems $\omega_{k,2^q}$ and $\Omega_{k,2^q}$.

Lemma 8 *For a canonical H_k^2,*

$$\begin{aligned}
\omega_{k,2^q}^L + \Omega_{k,2^q}^R &= (2^{2k}-1)\mathcal{N}_{k,2^q}^S, & \omega_{k,2^q}^R + \Omega_{k,2^q}^L &= (2^{2k}-1)\mathcal{N}_{k,2^q}^S, \\
\omega_{k,2^q}^T + \Omega_{k,2^q}^T &= (2^{2k}-1)\mathcal{N}_{k,2^q}^S, & \omega_{k,2^q}^B + \Omega_{k,2^q}^B &= (2^{2k}-1)\mathcal{N}_{k,2^q}^S.
\end{aligned}$$

3.3 Query Subgrids Overlapping with All Quadrants

For a $2^q \times 2^q$ query subgrid $G \subseteq \mathcal{R}$, we have: (1) $\theta_2(G) \in Q_4(H_k^2)$ and (2) $\theta_1(G) \in Q_1(H_k^2)$ by Remark 3.

For (1), when zooming in on the incomplete rectangular subgrid $G \cap Q_4(H_k^2)$ (with both side-lengths at most $2^q - 1$), we reduce $\sum_{\text{all } G \subseteq \mathcal{R}} \hbar_k(\theta_2(G), \partial_1(H_k^2))$ to $\Omega_{k-1,2^q}^{c_2}$ after $(+\frac{\pi}{2})$-rotating and reflecting $Q_4(H_k^2)$ into a canonical H_{k-1}^2 (with adjustment of distance cumulation).

For (2), similar consideration leads to a reduction of $\sum_{\text{all } G \subseteq \mathcal{R}} \hbar_k(\theta_1(G), \partial_1(H_k^2))$ to $\omega_{k-1,2^q}^{c_3}$, where $\omega_{k,2^q}^{c_3}$ denotes $\sum_{x=2^k-2^q+2}^{2^k} \sum_{y=2^k-2^q+2}^{2^k} \hbar_k(\theta_1(G_k(x,y,2^k,2^k)), \partial_1(H_k^2))$ for a canonical H_k^2 and is related to $\Omega_{k,2^q}^{c_2}$ as follows.

Lemma 9 *For a canonical H_k^2, $\omega_{k,2^q}^{c_3} + \Omega_{k,2^q}^{c_2} = (2^{2k}-1)\mathcal{N}_{k,2^q}^c$.*

Thus, the summation of all inter-cluster distances over all $2^q \times 2^q$ query subgrids contained in $\mathcal{R}$ is

$$(\Omega_{k-1,2^q}^{c_2} + 3 \cdot 2^{2k-2}\mathcal{N}_{k-1,2^q}^c) - \omega_{k-1,2^q}^{c_3} - (2^{2q}-1)\mathcal{N}_{k-1,2^q}^c.$$

3.4 The Big Picture: Computing $\Psi_q(H_k^2)$

The results in the previous three subsections yield $\varepsilon_{k,q}(H_k^2)$. Hence, we have the following recurrence for $\Psi_q(H_k^2)$:

$$\Psi_q(H_k^2) = \begin{cases} 4\Psi_q(H_{k-1}^2) + (\Omega^B_{k-1,2^q} + 2^{2k-2}\mathcal{N}^S_{k-1,2^q}) - \omega^R_{k-1,2^q} - (2^{2q}-1)\mathcal{N}^S_{k-1,2^q} \\ \quad + (\Omega^L_{k-1,2^q} + 2\cdot 2^{2k-2}\mathcal{N}^S_{k-1,2^q}) - (\omega^R_{k-1,2^q} + 2^{2k-2}\mathcal{N}^S_{k-1,2^q}) - (2^{2q}-1)\mathcal{N}^S_{k-1,2^q} \\ \quad + (\Omega^L_{k-1,2^q} + 3\cdot 2^{2k-2}\mathcal{N}^S_{k-1,2^q}) - (\omega^B_{k-1,2^q} + 2\cdot 2^{2k-2}\mathcal{N}^S_{k-1,2^q}) - (2^{2q}-1)\mathcal{N}^S_{k-1,2^q} \\ \quad + (\Omega^T_{k-1,2^q} + 3\cdot 2^{2k-2}\mathcal{N}^S_{k-1,2^q}) - (\omega^T_{k-1,2^q}) - (2^{2q}-1)\mathcal{N}^S_{k-1,2^q} \\ \quad + (\Omega^{c_2}_{k-1,2^q} + 3\cdot 2^{2k-2}\mathcal{N}^c_{k-1,2^q}) - (\omega^{c_3}_{k-1,2^q}) - (2^{2q}-1)\mathcal{N}^c_{k-1,2^q} & \text{if } k > q, \\ 0 & \text{if } k = q. \end{cases}$$

The exact formula for $\Psi_q(H_k^2)$ is:

$$\begin{aligned} \Psi_q(H_k^2) = {} & \frac{17}{14}\cdot 2^{3k+q} - \frac{17}{14}\cdot 2^{3k} - \frac{20885}{8151}\cdot 2^{2k+2q} + \frac{139}{48}\cdot 2^{2k+q} + \frac{7}{39}\cdot 2^{2k-2q}\cdot 3^q - 2^{2k-1} - \frac{1}{3}\cdot 2^{2k-q-2} \\ & + \frac{31}{1254}\cdot 2^{2k-2q}((\frac{3+\sqrt{5}}{2})^q + (\frac{3-\sqrt{5}}{2})^q) + \frac{21\cdot\sqrt{5}}{2090}\cdot 2^{2k-2q}((\frac{3+\sqrt{5}}{2})^q + (\frac{3-\sqrt{5}}{2})^q) \\ & + \frac{29767}{21736}\cdot 2^{k+3q} - 13\cdot 2^{k+2q-4} - 2^{k+q} - \frac{7}{39}\cdot 2^{k-q}\cdot 3^q + 3\cdot 2^{k-2} \\ & - \frac{63\cdot\sqrt{5}}{2090}\cdot 2^{k-q}((\frac{3+\sqrt{5}}{2})^q - (\frac{3-\sqrt{5}}{2})^q) - \frac{31}{418}\cdot 2^{k-q}((\frac{3+\sqrt{5}}{2})^q + (\frac{3-\sqrt{5}}{2})^q) - \frac{755}{35112}\cdot 2^{4q} \\ & - \frac{73}{21}\cdot 2^{3q-2} + 3\cdot 2^{2q-1} + \frac{21\cdot\sqrt{5}}{1045}((\frac{3+\sqrt{5}}{2})^q - (\frac{3-\sqrt{5}}{2})^q) + \frac{31}{627}((\frac{3+\sqrt{5}}{2})^q + (\frac{3-\sqrt{5}}{2})^q) - \frac{1}{3}\cdot 2^{q+1}. \end{aligned}$$

3.5 Total Number of Inter-cluster Gaps

In order to compute the (universe) mean inter-cluster distance over all inter-cluster gaps from all identically shaped subgrids, we need to derive the total number of inter-cluster gaps, denoted by $\Phi_{k,q}(H_k^2)$ for a canonical H_k^2.

For a grid space indexed by a space-filling curve, since the clusters interleave with the inter-cluster gaps of every query subgrid, we have:

$$\text{total number of inter-cluster gaps} = \text{total number of clusters} - \text{total number of query subgrids}.$$

As observed in [MJFS01],

$$\text{total number of clusters} = \text{total number of edges cut by all query subgrids} / 2.$$

They derive the exact formula for $E_{k,q}(H_k^2)$, which denotes the total number of edges cut by all $2^q \times 2^q$ query subgrids.

Alternatively, we follow a recursive approach similar to the computation of $\Psi_q(H_k^2)$, and develop a recurrence for $E_{k,q}(H_k^2)$:

$$E_{k,q}(H_k^2) \quad = \quad \begin{cases} 4E_{k,q}(H_{k-1}^2) + \Upsilon^R_{k-1,2^q} + \Upsilon^B_{k-1,2^q} + (2^{2q}-1) \\ \quad + \Upsilon^R_{k-1,2^q} + \Upsilon^L_{k-1,2^q} + \Upsilon^B_{k-1,2^q} + (2^{2q}-1) + \Upsilon^L_{k-1,2^q} \\ \quad + \Upsilon^T_{k-1,2^q} + \Upsilon^T_{k-1,2^q} + \Upsilon^{c_3}_{k-1,2^q} + \Upsilon^{c_4}_{k-1,2^q} + \Upsilon^{c_1}_{k-1,2^q} + \Upsilon^{c_2}_{k-1,2^q} & \text{if } k > q, \\ 2 & \text{if } k = q, \end{cases}$$

where Υ (replacing Ω) denotes the desired statistics (edges cut by query subgrid) over the four side-regions (L, R, B, and T) and four center-regions (c_1, c_2, c_3, and c_4).

The closed-form solution for $E_{k,q}(H_k^2)$ is

$$2^{2k+q+1} - 2^{k+2q+2} + 2^{k+q+1} + 2^{k-q+1} + 2^{3q+1} - 2^{2q+1}.$$

Hence,

$$\begin{aligned}\Phi_{k,q}(H_k^2) &= \frac{E_{k,q}(H_k^2)}{2} - (2^k - 2^q + 1)^2 \\ &= 2^{2k+q} - 2^{2k} - 2^{k+2q+1} + 3\cdot 2^{k+q} - 2^{k+1} + 2^{k-q} + 2^{3q} - 2^{2q+1} + 2^{q+1} - 1.\end{aligned}$$

4 Comparisons and Verification

By applying the same recursive approaches as in Section 3 to z-order curve family $\{Z_k^2 \mid k = 1,2,\ldots\}$, we obtain the summation of all inter-cluster distances over all $2^q \times 2^q$ query subgrids of a Z_k^2-structural grid space $[2^k]^2$,

$$\Psi_q(Z_k^2) = 2^{3k+q} - 2^{3k} - 2^{2k+2q+1} + 2^{2k+q+1} + 2^{k+3q} - 2^{k+q+1} + 2^k - 2^{3q} + 2^{2q+1} - 2^q$$

and the total number of inter-cluster gaps,

$$\Phi_{k,q}(Z_k^2) = 2^{2k+q+1} - 3\cdot 2^{2k} + 3\cdot 2^{2k-q-1} - 2^{2k-2q-1} - 2^{k+2q+2} + 2^{k+q+3} - 3\cdot 2^{k+1} + 2^{k-q+1} + 2^{3q+1} - 5\cdot 2^{2q} + 2^{q+2} - 1.$$

For a space-filling curve C_k indexing the grid space $[2^k]^2$, denote by $\Delta_{k,q}(C_k)$ the universe mean inter-cluster distance over all inter-cluster gaps from all $2^q \times 2^q$ subgrids of the C_k-structural grid space, and by $\widetilde{\Delta}_{k,q}(C_k)$ the mean total inter-cluster distance over all $2^q \times 2^q$ subgrids of the C_k-structural grid space.

The exact formulas for $\Psi_q(H_k^2)$, $\Phi_{k,q}(H_k^2)$, $\Psi_q(Z_k^2)$, and $\Phi_{k,q}(Z_k^2)$ give the exact formulas for $\Delta_{k,q}(H_k^2)$, $\Delta_{k,q}(Z_k^2)$, $\widetilde{\Delta}_{k,q}(H_k^2)$, and $\widetilde{\Delta}_{k,q}(Z_k^2)$. We simplify the exact results asymptotically as follows. For sufficiently large k and q with $k \gg q$ (typical scenario for range queries),

$$\Delta_{k,q}(C_k) \approx \begin{cases} \frac{17}{14}\cdot 2^k & \text{if } C_k \text{ is } H_k^2, \\ \frac{1}{2}\cdot 2^k & \text{if } C_k \text{ is } Z_k^2; \end{cases} \qquad \widetilde{\Delta}_{k,q}(C_k) \approx \begin{cases} \frac{17}{14}\cdot 2^{k+q} & \text{if } C_k \text{ is } H_k^2, \\ 2^{k+q} & \text{if } C_k \text{ is } Z_k^2; \end{cases}$$

$$\frac{\Delta_{k,q}(H_k^2)}{\Delta_{k,q}(Z_k^2)} \approx \frac{17}{7} \approx 2.43, \qquad \frac{\widetilde{\Delta}_{k,q}(H_k^2)}{\widetilde{\Delta}_{k,q}(Z_k^2)} \approx \frac{17}{14} \approx 1.21.$$

With respect to the $\Delta_{k,q}$-statistics, the z-order curve family clearly performs better than the Hilbert curve family over the considered ranges for k and q. With respect to the $\widetilde{\Delta}_{k,q}$-statistics, the superiority of z-order curve family persists but declines significantly.

We have verified all the exact formulas (intermediate and final) involved in the derivations in the analytical study with computer programs over various grid- and subgrid-orders: $k \in \{3,4,\ldots,10\}$ and $q \in \{2,3,\ldots,k\}$.

The random-walk model formulated provides good approximation to the true mean inter-clustering distance statistics for 2-dimensional order-k Hilbert curve H_k^2: the approximate statistics is asymptotically $\frac{2}{\sqrt{\pi}}2^k \approx 1.1284\cdot 2^k$ and the true statistics $\Delta_{k,q}(H_k^2) \approx 1.2143\cdot 2^k$.

5 Conclusion

We formulate a multi-dimensional random walk to study the inter-clustering performance of continuous multi-dimensional space-filling curves, and obtain a closed-form approximation to the universe mean inter-clustering distance for the Hilbert curve family. The excellent agreement suggests that the random walk may furnish an effective model to develop approximations to clustering and locality statistics for space-filling curves.

The principal random elements in our random-walk model depend solely on the edge-direction distribution of the corresponding space-filling curve. For general space-filling curves, such as the non-continuous z-order curve family, both spectra of rectilinear and non-rectilinear edge-direction distributions translate into the underlying transition probabilities. The added non-rectilinear transitions with non-unit step-size are expected to complicate required probabilistic analyses. However, more statistical/approximation applications of random walk for space-filling curves in general (dimensionality and continuity) settings, in which the edge-direction distribution and topological characteristics of the modeled space-filling curve are mathematically formulated into the principal random elements, are desired in order to confirm its robustness.

Our analytical study of the inter-clustering performances of 2-dimensional order-k Hilbert and z-order curve families are based upon the two inter-clustering statistics $\Delta_{k,q}$ and $\widetilde{\Delta}k,q$ — universe mean inter-cluster distance over all inter-cluster gaps and mean total inter-cluster distance over all subgrids of size $2^q \times 2^q$, respectively. The exact results allow us to compare their relative performances with respect to these two measures. For sufficiently large k and q with $k \gg q$, z-order curve family performs significantly (marginally) better than Hilbert curve family with respect to $\Delta_{k,q}$-statistics ($\widetilde{\Delta}_{k,q}$-statistics, respectively). We also verify the results with computer programs over various grid- and subgrid-orders.

A similar analytical study with dimensions greater than 3 appears to be much more difficult due to the loss of geometric intuition. The analysis of clustering properties of space-filling curves in [MJFS01] indicates that the Hilbert curve achieves better clustering than the z-order curve. Analytical studies that identify or characterize space-filling curve families that exhibit good clustering and inter-clustering performances simultaneously would be interesting.

References

[Alb97] J. Alber. Locality properties of discrete space-filling curves: Results with relevance for computer science (in German). Studienarbeit Universität Tübingen, Wilhelm-Schickard-Institut für Informatik. July 1997.

[AN00] J. Alber and R. Niedermeier. On multi-dimensional curves with Hilbert property. *Theory of Computing Systems*, 33(4):295–312, 2000.

[ARR+97] T. Asano, D. Ranjan, T. Roos, E. Welzl, and P. Widmayer. Space-filling curves and their use in the design of geometric data structures. *Theoretical Computer Science*, 181(1):3–15, 1997.

[BBK01] C. Böhm, S. Berchtold, and D. A. Keim. Searching in high-dimensional spaces — index structures for improving the performance of multimedia databases. *ACM Computing Surveys*, 33(3):322–373, 2001.

[BRWW97] E. Bugnion, T. Roos, R. Wattenhofer, and P. Widmayer. Space filling curves versus random walks. In van Kreveld, Nievergelt, Roos, and Widmayer, editors, *Lecture Notes in Computer Science (1340): Algorithmic Foundations of Geographic Information Systems*, pages 199–211, Springer-Verlag, Berlin Heidelberg, 1997.

[DS03] H. K. Dai and H. C. Su. On the locality properties of space-filling curves. To appear in *Proceedings of the 14th Annual International Symposium on Algorithms and Computation*, December 2003.

[GKP94] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics, Second Edition*. Addison-Wesley, Reading, Massachusetts, 1994.

[GL96] C. Gotsman and M. Lindenbaum. On the metric properties of discrete space-filling curves. *IEEE Transactions on Image Processing*, 5(5):794–797, 1996.

[Jag97] H. V. Jagadish. Analysis of the Hilbert curve for representing two-dimensional space. *Information Processing Letters*, 62(1):17–22, 1997.

[MD86] G. Mitchison and R. Durbin. Optimal numberings of an $N \times N$ array. *SIAM Journal on Algebraic and Discrete Methods*, 7(4):571–582, 1986.

[MJFS01] B. Moon, H. V. Jagadish, C. Faloutsos, and J. H. Saltz. Analysis of the clustering properties of the Hilbert space-filling curve. *IEEE Transactions on Knowledge and Data Engineering*, 13(1):124–141, 2001.

[NRS97] R. Niedermeier, K. Reinhardt, and P. Sanders. Towards optimal locality in mesh-indexings. In B. Chlebus and L. Czaja, editors, *Lecture Notes in Computer Science (1279): Fundamentals of Computation Theory, the Eleventh International Symposium*, pages 364–375, Springer-Verlag, Berlin Heidelberg, 1997.

[Sag94] H. Sagan. *Space-Filling Curves*. Springer-Verlag, New York, 1994.