Density of Ham- and Lee- non-isometric k-ary Words ⋆

Marcella Anselmo,[1] Manuela Flores,[2] Maria Madonia[3]

[1] Dipartimento di Informatica, Università di Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano (SA) Italy. E-mail: manselmo@unisa.it
[2] Dipartimento di Matematica e Informatica, Università di Palermo, Via Archirafi 34, 90123 Palermo, Italy. E-mail: manuela.flores@unipa.it
[3] Dipartimento di Matematica e Informatica, Università di Catania, Viale Andrea Doria 6/a, 95125 Catania, Italy. E-mail: madonia@dmi.unict.it

Abstract. Isometric k-ary words have been defined referring to the Hamming and the Lee distances. A word is non-isometric if and only if it has a prefix at distance 2 from the suffix of same length; such a prefix is called 2-error overlap. The limit density of isometric binary words based on the Hamming distance has been evaluated by Klavžar and Shpectorov, obtaining that about 8% of all binary words are isometric. In this paper, the issue is addressed for k-ary words and referring to the Hamming and the Lee distances. Actually, the only meaningful case of Lee-isometric k-ary words is when $k = 4$. It is proved that, when the length of words increases, the limit density of quaternary Ham-isometric words is around 17%, while the limit density of quaternary Lee-isometric words is even bigger, it is about 30%. The results are obtained using combinatorial methods and algorithms for counting the number of k-ary isometric words.

Keywords: Isometric words, Overlap with errors, Hamming and Lee distance, Density

1 Introduction

The notion of isometric word has been introduced in the framework of the research on hypercubes and, more in general, on k-ary n-cubes. The k-ary n-cube is one of the most attractive interconnection networks for parallel computer systems. The goal was to provide a class of subgraphs of the hypercube Q_n having a considerably smaller size, still maintaining some metric properties. With this aim, Hsu introduced the Fibonacci cubes [12], as the subgraphs of the hypercube restricted to vertices associated with binary words that do not contain 11 as a factor. Fibonacci cubes are isometric subgraphs of Q_n. They received a lot of attention afterwards (see [14] for a survey) and they have been then extended to define the generalized Fibonacci cube $Q_n(f)$ [13], as the subgraph of the hypercube Q_n restricted to the vertices associated with binary words avoiding the word f as a factor, i.e. *f-free* binary words. In this framework, a binary word f is said

⋆ Partially supported by INdAM-GNCS Project 2022, FARB Project ORSA229894 of University of Salerno, PNRR MUR Project ITSERR CUP B53C22001770006 and FFR fund University of Palermo, TEAMS Project and PNRR MUR Project PE0000013-FAIR University of Catania.

isometric when, for any $n \geq 1$, $Q_n(f)$ can be isometrically embedded into Q_n, and *non-isometric*, otherwise [15]. For example, the word 11 is isometric, because Fibonacci cubes $Q_n(11)$ are isometric subgraphs of Q_n. Other examples are given in Examples 1, 5 and 7.

Observe that, in the binary case, the distance between two vertices in the hypercube coincides with their *Hamming distance*. Hence, isometric binary words can be characterized ignoring hypercubes and adopting a point of view closer to combinatorics on words. A binary word f is isometric if and only if for any integer $d \geq |f|$ and any pair of words u and v of length d which do not contain the factor f, u can be transformed in v by exchanging one by one the bits on which they differ and generating only words which do not contain f. Differently saying, this transformation is composed by single steps transforming a word in another at Hamming distance 1. We will call it an f-free *Ham-transformation* and the resulting isometric words, *Ham-isometric words*. When moving from binary to k-ary alphabets, with $k \geq 2$, the hypercubes are replaced by the k-ary n-cubes where the vertices are k-ary words of length n. In this case, the distance between two vertices is no more captured by the Hamming distance, but by the *Lee distance*. Hence, in an analogous way, f-free *Lee-transformations* and *Lee-isometric k-ary words* have been introduced; see [5], and [3] on quaternary words. Remarkably, note that Lee-isometric words exist only for k-ary alphabets with $k = 2,3,4$, whereas there are Ham-isometric words for any cardinality of the alphabet. Further note that when $k = 2,3$ the two notions coincide, so that the unique meaningful case to investigate Lee-isometric words is when the alphabet is quaternary.

The notion of isometric word combines the distance notion with the property that a word does not appear as factor in other words. Note that this property is important in combinatorics as well as in the investigation on similarities, or distances, on DNA sequences, where the avoided factor is referred to as an absent or forbidden word [8–11]. Recently, isometric words have been introduced and investigated in [1, 2] referring to an edit distance based on swap and mismatch errors. Also, binary non-isometric words have been considered in the two-dimensional setting, and non-isometric/bad pictures have been investigated [6].

Deciding whether a word is Ham-isometric (Lee-isometric, resp.) can be efficiently done using the characterization of Ham-non-isometric (Lee-non-isometric, resp.) words as the ones showing a particular overlap with errors, called 2-Ham-error overlap (2-Lee-error overlap, resp.). A 2-*Ham-error overlap* (2-*Lee-error overlap*, resp.) of a word f is a prefix of f whose Hamming (Lee, resp.) distance from the suffix of same length is exactly 2. This is a similar concept as the overlap, or border, of a word, i.e. a prefix which is equal to the suffix of same length. Words having no overlap are known in the literature as the non bifix-free words or unbordered words. Such notions play a crucial role both in combinatorics of words and in pattern matching (with or without errors).

In [15], the authors demonstrate that there is a considerable number of both Ham-isometric and Ham-non-isometric binary words. In fact, they show that, as the length goes to infinity, the proportion of Ham-isometric words has a limit strictly between 0 and 1. The density of the set of all binary words of given length having a 2-error overlap converges to a limit value which lies between 0.919975 and 0.924156, that is there are about the 8% of Ham-isometric binary words. Thus, the generalized Fibonacci cubes

$Q_n(f)$ for Ham-isometric binary words f constitute a large explicit family of partial cubes. Actually, the evaluation of the density of Ham-isometric binary words has been achieved using their characterization as those words without 2-error overlaps.

In this paper we extend such results by proving that Ham- and Lee- isometric words over a k-ary alphabet, with $k > 2$, can be even more than in the binary case. The density of Ham- isometric k-ary words is investigated for any k; upper and lower bounds are given depending on k and on the length n of words for which the density can be explicitely computed. Here, the computation has been carried on for $k = 4$ and $n = 3, ..., 16$, and the values are collected in a table. In an analogous way, the density of Lee-isometric words has been lower and upper bounded. Recall that there are no Lee-isometric words for $k > 5$, and that Lee-isometric words are exactly the Ham-isometric words, when $k = 2, 3$. So the results concern the unique meaningful case of $k = 4$. In the quaternary case, the density of Ham-isometric and Lee-isometric words has been explicitely evaluated and compared. There are about the 17% of Ham-isometric quaternary words, whereas about 30% of Lee-isometric quaternary words. Remarkably, there are strictly more Lee-isometric quaternary words than the Ham- ones. The motivation of this claim has been explored.

The computation of explicit values of the density of Ham- and Lee- isometric quaternary words for small lenghts has been carried on using an algorithm to efficiently check whether a word is isometric. A first cubic time algorithm for deciding isometricity and providing evidence and further information about it was given in [16] for binary words and referring to the Hamming distance. Recently, an algorithm has been presented to check isometricity of k-ary words with Hamming and Lee distances [7]. This algorithm is based on the characterization in [3] and applies some methods of the pattern matching with mismatches to achieve a linear time complexity. Note that, from then on, other algorithms have been designed that, not only check whether a k-ary word is Ham- or Lee- isometric, but they also provide further information and evidence while keeping the same linear complexity [4].

2 Isometric Words and 2-error overlaps

Let us recall some definitions and notation given in [5].

Let Σ be an alphabet and $|\Sigma| = k$. Throughout the paper, Σ will be identified with $\mathbb{Z}_k = \{0, 1, \ldots, k-1\}$, the ring of integers modulo k. A word (or string) $f \in \Sigma^*$ of length n is $f = x_1 x_2 \cdots x_n$, where $x_1, x_2, \ldots, x_n$ are symbols in Σ. The set of words over Σ of length n is denoted Σ^n. Let $f[i]$ denote the symbol of f in position i, i.e. $f[i] = x_i$. Then, $f[i..j] = x_i \cdots x_j$, for $1 \leq i \leq j \leq n$, is a factor of f. A word $s \in \Sigma^*$ is said f-free if it does not contain f as a factor. The prefix of f of length l is $pre_l(f) = f[1..l]$; while the suffix of f of length l is $suf_l(f) = f[n-l+1..n]$. When $pre_l(f) = suf_l(f)$ then $pre_l(f)$ is referred to as an overlap, or border, of f of length l.

Let $u, v \in \Sigma^*$ be two words of the same length. The *Hamming distance* $dist_H(u, v)$ between u and v is the number of positions at which u and v differ.

The *Lee distance* between two words $u, v \in \mathbb{Z}_k^n$, $u = x_1 \cdots x_n$ and $v = y_1 \cdots y_n$ is $dist_L(u, v) = \sum_{i=1}^{n} min(|x_i - y_i|, k - |x_i - y_i|)$.

In the sequel, Σ will denote a generic alphabet of cardinality k, while Δ denote the quaternary alphabet $\Delta = \{A, C, T, G\}$, referred to as the *genetic alphabet*. Symbols A and T (C and G, resp.) will be called *complementary symbols*, in analogy to the Watson-Crick complementary bases they represent. The alphabet Δ will be identified with $\mathbb{Z}_4$, in such a way that A, C, T, and G will be identified with 0, 1, 2, and 3, respectively. Therefore, pairs of complementary symbols have Lee distance 2, whereas pairs of distinct non-complementary symbols have Lee distance 1.

Let us now recall the definitions of Ham and Lee-isometric words [5]. The definitions are based on the process of transforming a word into another one of equal length, changing one symbol at a time. Let Σ be a k-ary alphabet, $f \in \Sigma^n$, and $u, v \in \Sigma^d$.
A *Ham-transformation* (*Lee-transformation*, resp.) of length h from u to v is a sequence of words $w_0, w_1, \ldots, w_h$ such that $w_0 = u$, $w_h = v$, and for any $i = 0, 1, \ldots, h-1$, $\mathrm{dist}_H(w_i, w_{i+1}) = 1$ ($\mathrm{dist}_L(w_i, w_{i+1}) = 1$, resp.). If for any $i = 0, 1, \ldots, h$, the word w_i is f-free, then the Ham-transformation (Lee-transformation, resp.) is said *f-free*.

A word $f \in \Sigma^n$ is *Ham-isometric* (*Lee-isometric*, resp.) if for all $d \geq n$, and f-free words $u, v \in \Sigma^d$, there is an f-free Ham-transformation (Lee-transformation, resp.) from u to v of length equal to $dist_H(u,v)$ ($dist_L(u,v)$, resp.). A word is *Ham-non-isometric* (*Lee-non-isometric*, resp.) if it is not Ham-isometric (Lee-isometric, resp.).

A pair (u,v) of words u, $v \in \Sigma^d$ is referred to as a pair of *Ham-witnesses*, (*Lee-witnesses*, resp.) for a Ham- (Lee- , resp.) non-isometric word f, if u and v are f-free words and there does not exist an f-free Ham- (Lee-, resp) transformation from u to v of length equal to $dist_H(u,v)$ ($dist_L(u,v)$, resp.).

Example 1. Let Δ be the quaternary genetic alphabet, $f = ACT$, $u = ACCCT$, and $v = ACGCT$. Observe that $dist_L(u,v) = 2$, since they differ in their third position only and $dist_L(C,G) = 2$. The sequences $ACCCT, ACACT, ACGCT$ and $ACCCT, ACTCT, ACGCT$ are two Lee-transformations from u to v of length equal to $dist_L(u,v) = 2$; they are not f-free. Actually, no f-free Lee-transformation exists from u to v. This shows that ACT is Lee-non-isometric and that (u,v) is a pair of Lee-witnesses for ACT.

Let us recall the following definitions (see Figure 1) of Ham- and Lee-error overlap.

Definition 2. *Let Σ be a k-ary alphabet, $f \in \Sigma^n$, and q be an integer, $1 \leq q \leq n-1$.*
*The word f has a q-*Ham-error overlap *(q-*Lee-error overlap*, resp.) of length l, $1 \leq l \leq n-1$, if $dist_H(pre_l(f), suf_l(f)) = q$ ($dist_L(pre_l(f), suf_l(f)) = q$, resp.). Its* error positions *are the q (m, $1 \leq m \leq q$, resp.) positions in $pre_l(f)$ where it differs from $suf_l(f)$.*

Remark 3. Using the notations in the previous definition, if f has a q-Lee-error overlap of length l, then $1 \leq m \leq l, q$.
In particular, when $k = 4$ and $q = 2$, then $m = 1$ or $m = 2$. The case $m = 1$ holds if $pre_l(f)$ and $suf_l(f)$ differ in exactly one position and the error is given by a pair of complementary symbols. For example, $f = AGAC \in \Delta^4$ has a 2-Lee-error overlap of length $l = 2$. Indeed, $m = 1$ and $dist_L(AG, AC) = 2$. If $m = 2$ then $pre_l(f)$ and $suf_l(f)$ differ in two different positions i and j and the errors are given by pairs of non-complementary symbols.

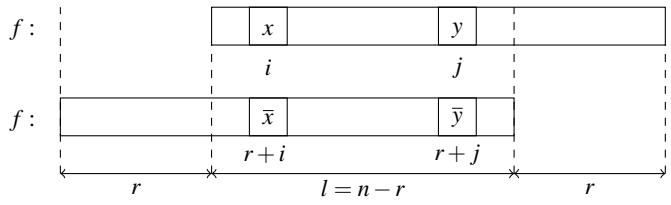

Fig. 1. The word f and its 2-error overlap of length l

Theorem 4, proved in [3, 5], provides a characterization of Ham- and Lee- isometric words, which is fundamental to test whether a word is Ham- or Lee- isometric.

Theorem 4 ([3, 5]). *Let Σ be a k-ary alphabet and $f \in \Sigma^*$. Then,*

- *f is Ham-isometric if and only if it has no 2-Ham-error overlap.*
- *f is Lee-isometric if and only if it has no 2-Lee-error overlap, when $k = 2,3,4$*
- *f is never Lee-isometric, when $k > 4$.*

Example 5. Let $f = 0201 \in \Sigma^*$ with $\Sigma = \mathbb{Z}_3 = \{0,1,2\}$. The word f has no 2-Ham-error overlap and thus it is Ham-isometric, by Theorem 4. Consider now $f = ATC \in \Delta^*$. The word f has no 2-Lee-error overlap and thus it is Lee-isometric, by Theorem 4. On the other hand, by the same theorem, $f = ATC$ is Ham-non-isometric, since it has a 2-Ham-error overlap.

Next result allows us to restrict the domain of strings to be considered when looking for Lee-isometric words. For example, when the alphabet is Δ, it is sufficient to take into account words starting with A.

Let $\Sigma = \{0,1,\ldots,k-1\}$, $f = f_1 f_2 \cdots f_n$ be a word over Σ, $u, v \in \Sigma^d$ be f-free words, and $h, j \in \Sigma$. The *reverse* of f is $f^R = f_n \cdots f_2 f_1$. The *h-shift* of j is $j^{S(h)} = (j+h) \bmod k$, while the h-shift of f is $f^{S(h)} = f_1^{S(h)} f_2^{S(h)} \cdots f_n^{S(h)}$. When $k = 2$, the 1-shift of f is its complement.

Lemma 6. *Let Σ be a k-ary alphabet and $f \in \Sigma^*$. Then*

- *f is Lee-isometric if and only if f^R is Lee- isometric*
- *for any $h \in \Sigma$, f is Lee-isometric if and only if $f^{S(h)}$ is Lee-isometric.*

3 Evaluating the Density of Ham- and Lee- isometric Words

The density of Ham-isometric binary words has been studied in [15], where the authors show that, for large values of the length, about 8% of all binary words are Ham-isometric. In this section, the case of an alphabet with k symbols, $k \geq 2$, is investigated. Results concern both Ham- and Lee- isometric words and will be obtained using their characterizations in terms of 2-error overlaps (see Theorem 4).

3.1 Density of Ham-isometric words

Let us evaluate the density of Ham-non-isometric words, i.e., words with a 2-Ham-error overlap, as the length increases. Table 1 collects the values of the density of quaternary Ham-non-isometric words of length n, with $3 \leq n \leq 16$.

n	h_n	$\widehat{h_n}$	α_n	$\widehat{\alpha_n}$
3	36	24	0,5625	0,375
4	168	152	0,65625	0,59375
5	804	624	0,78515625	0,609375
6	3228	2704	0,788085938	0,66015625
7	13404	11176	0,818115234	0,682128906
8	54516	45360	0,831848145	0,692138672
9	216756	183656	0,826858521	0,700592041
10	875052	737008	0,834514618	0,702865601
11	3490236	2956520	0,832137108	0,704889297
12	13994460	11828800	0,834134817	0,705051422
13	55909620	47356176	0,83311826	0,705662012
14	223809540	189392808	0,833755508	0,70554319
15	894723276	757694840	0,833275985	0,705658309
16	3579796572	3030588552	0,83348634	0,705613883

Table 1. Some values of h_n, $\widehat{h_n}$, α_n, $\widehat{\alpha_n}$ for $n = 3, \ldots, 16$

The values in the table show that the density is not a monotone sequence. That is why we will separately consider the density of words with a "long" 2-error overlap, and of words with a "short" 2-error overlap.

Let $\mathcal{H}_{k,n}$ be the set of all k-ary words of length n having a 2-Ham-error overlap. Let $\mathcal{H}_{k,n}^{short}$ be the set of all words in $\mathcal{H}_{k,n}$ which have a 2-Ham-error overlap of length $l \leq n/2$, $\mathcal{H}_{k,n}^{long}$ be the set of all words in $\mathcal{H}_{k,n}$ which have a 2-Ham-error overlap of length $l > n/2$. A word in $\mathcal{H}_{k,n}^{short}$ is called *split* (as in [15], for $k = 2$).

Clearly, $\mathcal{H}_{k,n} = \mathcal{H}_{k,n}^{short} \cup \mathcal{H}_{k,n}^{long}$, but $\mathcal{H}_{k,n}^{short} \cap \mathcal{H}_{k,n}^{long}$ is not necessarily empty. In particular, $|\mathcal{H}_{k,n}| \leq |\mathcal{H}_{k,n}^{short}| + |\mathcal{H}_{k,n}^{long}|$. Also note that $\mathcal{H}_{k,n} \setminus \mathcal{H}_{k,n}^{short} \subseteq \mathcal{H}_{k,n}^{long}$.

Example 7. Let $\Delta = \{A, C, T, G\}$ be the genetic alphabet and $f = AAGATAA$ in Δ^7. The word f is Ham-non-isometric. It has a 2-Ham-error overlap of length $l = 3$ that involves error positions $i = 1$ and $j = 3$ with $dist_H(AAG, TAA) = 2$. Since $l \leq n/2$, then $f \in \mathcal{H}_{4,7}^{short}$. Furthermore, f has also a 2-Ham-error overlap of length $l = 4$ that involves error positions $i = 2$ and $j = 3$ with $dist_H(AAGA, ATAA) = 2$. Since $l > n/2$, then $f \in \mathcal{H}_{4,7}^{long}$. Therefore, f belongs to both sets $\mathcal{H}_{4,7}^{short}$ and $\mathcal{H}_{4,7}^{long}$.

Let us denote $h_{k,n} = |\mathcal{H}_{k,n}|$, $s_{k,n} = |\mathcal{H}_{k,n}^{short}|$ and $l_{k,n} = |\mathcal{H}_{k,n}^{long}|$. From $|\mathcal{H}_{k,n}| \leq |\mathcal{H}_{k,n}^{short}| + |\mathcal{H}_{k,n}^{long}|$, it follows $h_{k,n} \leq s_{k,n} + l_{k,n}$. Further denote by $\alpha_{k,n}$, $\sigma_{k,n}$, and $\lambda_{k,n}$ the density of words in $\mathcal{H}_{k,n}$, $\mathcal{H}_{k,n}^{short}$, and $\mathcal{H}_{k,n}^{long}$, respectively, among all words of length n, i.e., $\alpha_{k,n} = \dfrac{h_{k,n}}{k^n}$, $\sigma_{k,n} = \dfrac{s_{k,n}}{k^n}$, and $\lambda_{k,n} = \dfrac{l_{k,n}}{k^n}$.

Let us start by counting the number of words with a 2-Ham-error-overlap of fixed length. Denote by $h_{k,n}(d)$ the number of words in $\mathcal{H}_{k,n}$ that have a 2-Ham-error overlap of length d, for some $2 \leq d \leq n-1$; by $s_{k,n}(d)$ the number of words in $\mathcal{H}_{k,n}^{short}$ that have a 2-Ham-error overlap of length d, for some $2 \leq d \leq \lfloor n/2 \rfloor$; and by $l_{k,n}(d)$ the number of words in $\mathcal{H}_{k,n}^{long}$ that have a 2-Ham-error overlap of length d, for some $\lfloor n/2 \rfloor < d \leq n-1$.

Lemma 8. *Let Σ be a k-ary alphabet. Then, $h_{k,n}(d) = \frac{d(d-1)}{2}(k-1)^2 k^{n-d}$.*

Proof. Let f be a k-ary word of length n that has a 2-error overlap of length d, for some $2 \leq d \leq n-1$. The word f is fully specified by three informations: the bits in the last $n-d$ positions of f, the 2 locations of the "errors" within $pref_d(f)$ and by the symbols in these error positions, which can be chosen in $k-1$ ways each. The number of choices of 2 positions within the d positions in $pref_d(f)$ is $\binom{d}{2}$. Hence, $h_{k,n}(d) = k^{n-d}\binom{d}{2}(k-1)^2 = \frac{d(d-1)}{2}(k-1)^2 k^{n-d}$. □

Remark 9. Note that a k-ary word f of length n may have 2-Ham-error overlaps of different lengths. This implies that $|\mathcal{H}_{k,n}| = h_{k,n} \leq \sum_{d=2}^{n-1} h_{k,n}(d)$. Similar reasonings show that $|\mathcal{H}_{k,n}^{short}| = s_{k,n} \leq \sum_{d=2}^{\lfloor n/2 \rfloor} s_{k,n}(d)$ and that $|\mathcal{H}_{k,n}^{long}| = l_{k,n} \leq \sum_{d=\lfloor n/2 \rfloor + 1}^{n-1} l_{k,n}(d)$.

Proposition 10. *Let Σ be a k-ary alphabet. The density of words in $\mathcal{H}_{k,n}^{long}$ converges to 0 as n goes to infinity. That is*

$$\lim_{n\to\infty} \lambda_{k,n} = 0.$$

Proof. Consider $l_{k,n}(d)$, the number of k-ary words that have a 2-error overlap of length exactly d, for some $\lfloor n/2 \rfloor < d \leq n-1$. From Lemma 8 and Remark 9, we have

$$l_{k,n} \leq \sum_{d=\lfloor n/2 \rfloor + 1}^{n-1} k^{n-d}\binom{d}{2}(k-1)^2 = (k-1)^2 \sum_{d=\lfloor n/2 \rfloor + 1}^{n-1} k^{n-d} d(d-1)/2.$$

Then

$l_{k,n} \leq (k-1)^2 \dfrac{k^{n/2}}{2} \sum_{d=\lfloor n/2 \rfloor + 1}^{n-1} d^2 \leq (k-1)^2 \dfrac{k^{n/2}}{2} n(n-1)^2$. Therefore,

$\lambda_{k,n} \leq \dfrac{(k-1)^2 k^{n/2} n(n-1)^2}{2k^n} = \dfrac{(k-1)^2 n(n-1)^2}{2k^{n/2}}$ and $\lim_{n\to\infty} \lambda_{k,n} = 0$. □

Contrarily to the case of the sequence $\alpha_{k,n}$, the following result holds.

Proposition 11. *Let Σ be a k-ary alphabet. The sequence $\sigma_{k,n}$ is monotonically increasing and bounded from above by 1. In particular, it has a limit $\sigma_k \leq 1$.*

Proof. Let us show that for any $n \geq 1$, $s_{k,n+1} \geq k s_{k,n}$ so that $\sigma_{k,n+1} = \frac{s_{k,n+1}}{k^{n+1}} \geq \frac{s_{k,n}}{k^n} = \sigma_{k,n}$. Consider the mapping $\varphi : \Sigma^{n+1} \to \Sigma^n$ defined by erasing the bit in position $\left\lfloor \frac{n}{2} \right\rfloor + 1$. Now, if $f \in \Sigma^{n+1}$ and $\varphi(f)$ has a 2-error overlap of some length $d \leq \left\lfloor \frac{n}{2} \right\rfloor$, then f has also a 2-error overlap of the same length d. Therefore, $\varphi^{-1}(\mathcal{H}^{short}_{k,n}) \subseteq \mathcal{H}^{short}_{k,n+1}$ and the claim follows noting that every $f \in \mathcal{H}^{short}_{k,n}$ is the image of k different elements in $\mathcal{H}^{short}_{k,n+1}$. □

Proposition 12. *Let Σ be a k-ary alphabet. The sequence $\alpha_{k,n}$ converges to the same limit value σ_k, as $\sigma_{k,n}$.*

Proof. According to Proposition 10, the sequence $\lambda_{k,n}$ tends to zero. Hence both $\sigma_{k,n}$ and $\sigma_{k,n} + \lambda_{k,n}$ converge to the same limit, σ_k. On the other hand, clearly, $\sigma_{k,n} \leq \alpha_{k,n} \leq \sigma_{k,n} + \lambda_{k,n}$, since $\mathcal{H}^{short}_{k,n} \subseteq \mathcal{H}_{k,n} \subseteq \mathcal{H}^{short}_{k,n} \cup \mathcal{H}^{long}_{k,n}$. So the claim follows. □

Let us estimate σ_k, the limit value of both density sequences $\sigma_{k,n}$ and $\alpha_{k,n}$.

Theorem 13. *Let Σ be a k-ary alphabet. The value σ_k of the limit density of Ham-non-isometric k-ary words is*

$$\sigma_{k,2m} \leq \sigma_k \leq \sigma_{k,2m} + f(k,m)$$

where, for any integer $m \geq 1$,

$$f(k,m) = \sum_{i=m}^{\infty} \frac{i(i+1)}{2k^{i-1}} = \frac{m^2k^2 - (2m^2 - 3m - 1)k + (m^2 - 3m + 2)}{2(k-1)^3 k^{m-2}}.$$

Proof. Using Proposition 11, the sequence $\sigma_{k,n}$ is monotonically increasing, hence, for any $n \geq 2$, $\sigma_{k,n} \leq \sigma_k$. Furthermore, $s_{k,2m+1} = k s_{k,2m}$ and then $\sigma_{k,2m+1} = \sigma_{k,2m}$, so that we only need to consider even $n = 2m$.

Let $t_{k,n}$ be the number of non-split words of length n, i.e., $t_{k,n} = |\Sigma^n \setminus \mathcal{H}^{short}_{k,n}|$. If w is such a word then inserting two new symbols in the middle produces a word of length $n+2$ which is either again non-split or it has a 2-error overlap of length exactly $m+1$. The number of words of the latter sort is $k^{m+1}\binom{m+1}{2}(k-1)^2$, because we can choose $m+1$ symbols arbitrarily and then the second half must be the same as the first half but with two positions changed in $k-1$ ways each. Therefore, $k^2 t_{k,n} \leq t_{k,n+2} + k^{m+1}\binom{m+1}{2}(k-1)^2$ and, dividing by k^{n+2}, $\frac{t_{k,n}}{k^n} \leq \frac{t_{k,n+2}}{k^{n+2}} + \frac{(k-1)^2 m(m+1)}{2k^{m+1}}$.

Referring to the densities $\mu_{k,n} = \frac{t_{k,n}}{k^n}$ of non-split k-ary words of length n, one has $\mu_{k,n+2} \geq \mu_{k,n} - \frac{(k-1)^2 m(m+1)}{2k^{m+1}}$.

Since $\sigma_{k,n} = 1 - \mu_{k,n}$, we get $\sigma_{k,n+2} \leq \sigma_{k,n} + \frac{(k-1)^2 m(m+1)}{2k^{m+1}} \leq \sigma_{k,n} + \frac{m(m+1)}{2k^{m-1}}$.

Combining these relations from n to $n+p$, we obtain

$$\sigma_{k,n+2p} \leq \sigma_{k,n} + \sum_{i=m}^{m+p-1} \frac{i(i+1)}{2k^{i-1}}.$$

Therefore, the following upper bound for σ_k follows: $\sigma_k \leq \sigma_{k,n} + \sum_{i=m}^{\infty} \frac{i(i+1)}{2k^{i-1}}$. Hence, for any $n \geq 2$, $n = 2m$, $\sigma_{k,n} \leq \sigma_k \leq \sigma_{k,n} + \sum_{i=m}^{\infty} \frac{i(i+1)}{2k^{i-1}}$.

Let us now evaluate $\sum_{i=m}^{\infty} \frac{i(i+1)}{2k^{i-1}}$. Note that $\sum_{i=m}^{\infty} \frac{i(i+1)}{2k^{i-1}} = \frac{k}{2} \sum_{i=m}^{\infty} (\frac{i^2}{k^i} + \frac{i}{k^i})$.

Setting $x = 1/k$ in classical formulas for $\sum_{i=1}^{\infty} i^2 x^i$, $\sum_{i=1}^{m-1} i^2 x^i$, $\sum_{i=1}^{\infty} i x^i$, and $\sum_{i=1}^{m-1} i x^i$, the following evaluation can be obtained

$$\sum_{i=m}^{\infty} \frac{i(i+1)}{2k^{i-1}} = \frac{(m^2+m)k^2 - (2m^2-2)k + (m^2-m)}{2(k-1)^3 k^{m-2}}. \qquad \square$$

3.2 Density of Lee-isometric words

In order to evaluate the density of Lee-isometric quaternary words some of the results regarding Ham-isometric words must be properly modified.

Remember that the only significant case is now the case of a quaternary alphabet. Let $\Delta = \{A, C, T, G\}$ be the alphabet with $k = |\Delta| = 4$. The main difference is that a word $f \in \Delta^*$ has a 2-Lee-error-overlap when, for some $d \leq n-1$, $pref_d(f)$ differs from $suf_d(f)$ in either 2 positions which contain different non-complementary symbols or 1 position which contains two complementary symbols.

In analogy to the case of the Hamming distance, let us state the following notations. Note that the value $k = 4$ is understood. Let $\mathcal{L}_n$ be the set of all words in Δ^* of length n having a 2-Lee-error overlap.

Let $\mathcal{L}_n^{short}$ be the set of all words in $\mathcal{L}_n$ which have a 2-Lee-error overlap of length $l \leq n/2$, while $\mathcal{L}_n^{long}$ be the set of all words in $\mathcal{L}_n$ which have a 2-Lee-error overlap of length $l > n/2$. A word in $\mathcal{L}_{k,n}^{short}$ is called *L-split*. Let us denote $\widehat{h}_n = |\mathcal{L}_n|$, $\widehat{s}_n = |\mathcal{L}_n^{short}|$, $\widehat{l}_n = |\mathcal{L}_n^{long}|$, and by $\widehat{\alpha}_n$, $\widehat{\sigma}_n$, and $\widehat{\lambda}_n$ the density of words in $\mathcal{L}_n$, $\mathcal{L}_n^{short}$, and $\mathcal{L}_n^{long}$, respectively, among all words of length n, i.e., $\widehat{\alpha}_n = \frac{\widehat{h}_n}{k^n}$, $\widehat{\sigma}_n = \frac{\widehat{s}_n}{k^n}$, and $\widehat{\lambda}_n = \frac{\widehat{l}_n}{k^n}$.

Finally, let $\widehat{h}_n(d)$ be the number of words in $\mathcal{L}_n$ that have a 2-Lee-error overlap of length d, for some $2 \leq d \leq n-1$; $\widehat{s}_n(d)$ be the number of words in $\mathcal{L}_n^{short}$ that have a 2-Lee-error overlap of length d, for some $2 \leq d \leq \lfloor n/2 \rfloor$; and $\widehat{l}_n(d)$ be the number of words in $\mathcal{L}_n^{long}$ that have a 2-Lee-error overlap of length d, for some $\lfloor n/2 \rfloor < d \leq n-1$.

Lemma 14. *The number of words in Δ^* of length n that have a 2-Lee-error overlap of length d, is $\widehat{h}_n(d) = \big(2d^2 - d\big)4^{n-d}$.*

Proof. Let f be a word in Δ^n that has a 2-Lee-error overlap of length d, for some $1 \leq d \leq n-1$. Then $pref_d(f)$ and $suf_d(f)$ differ either in two positions, and the errors are given by a pair of non-complementary symbols, or in one position, and the error is given by a pair of complementary symbols. Therefore, in the first case, f is fully specified by three informations: the bits in the last $n-d$ positions of f, the 2 locations

of the errors within $pref_d(f)$ and by the pairs of symbols in these error positions, which can be chosen in 4 different ways. In the second case, the word f is fully specified by two informations: the bits in the last $n-d$ positions of f and the location of the error within $pref_d(f)$. Hence, $\widehat{h}_n(d) = 4^{n-d}\left[4\binom{d}{2} + d\right] = \left(2d^2 - d\right)4^{n-d}$. □

Proposition 15. *The density of words in $\mathcal{L}_n^{long}$ converges to 0 as n goes to infinity, i.e.*

$$\lim_{n\to\infty} \widehat{\lambda}_n = 0.$$

Proof. Let $\widehat{l}_n(d)$ be the number of words $\in \Delta^n$ that have a 2-Lee-error overlap of length exactly d, for some $\lfloor n/2 \rfloor < d \leq n-1$. From Lemma 14, we have

$\widehat{l}_n \leq \sum_{d=\lfloor n/2 \rfloor+1}^{n-1} 4^{n-d}\left(2d^2 - d\right)$. Then,

$$\widehat{l}_n \leq 2 \cdot 4^{n/2} \sum_{d=\lfloor n/2 \rfloor+1}^{n-1} d^2 \leq 2 \cdot 4^{n/2} \sum_{d=\lfloor n/2 \rfloor+1}^{n-1} (n-1)^2 \leq 4^{n/2} n (n-1)^2.$$

Therefore, $\widehat{\lambda}_n = \dfrac{\widehat{l}_n}{4^n} \leq \dfrac{n(n-1)^2}{4^{n/2}}$ and $\lim_{n\to\infty} \widehat{\lambda}_n = 0$. □

The following propositions can be proved similarly to Propositions 11 and 12.

Proposition 16. *The sequence $\widehat{\sigma}_n$ is monotonically increasing and bounded from above by 1. In particular, it has a limit $\widehat{\sigma} \leq 1$.*

Proposition 17. *The sequence $\widehat{\alpha}_n$ converges to the same limit value $\widehat{\sigma}$, as $\widehat{\sigma}_n$.*

Let us estimate the limit density $\widehat{\sigma}$ of both sequences $\widehat{\sigma}_n$ and $\widehat{\alpha}_n$.

Theorem 18. *The limit value $\widehat{\sigma}$ of the density of Lee-non-isometric words in Δ^* is*

$$\widehat{\sigma}_{2m} \leq \widehat{\sigma} \leq \widehat{\sigma}_{2m} + f(m)$$

where, for any integer $m \geq 1$,

$$f(m) = \sum_{i=m}^{\infty} \frac{2i^2 + 3i + 1}{4^{i+1}} = \frac{18m^2 + 39m + 28}{27 \cdot 4^m}.$$

Proof. Using Proposition 16, the sequence $\widehat{\sigma}_n$ is monotonically increasing, hence, for any $n \geq 2$, $\widehat{\sigma}_n \leq \widehat{\sigma}$. Furthermore, $s_{2m+1} = 4s_{2m}$ and then $\widehat{\sigma}_{2m+1} = \widehat{\sigma}_{2m}$, so that we only need to consider even $n = 2m$.

Let $\widehat{t}_n$ be the number of non-L-split words of length n. If w is such a word then inserting two new symbols in the middle produces a word of length $n+2$ which is either again non-L-split or it has a 2-Lee-error overlap of length exactly $m+1$. The number of words of the latter sort is $4^{m+1}\left[4\binom{m+1}{2} + m + 1\right]$, because we can choose $m+1$ symbols arbitrarily and, then, the second half must be the same as the first half but either with two positions changed in 2 ways each (if the errors are given by a pair of non-complementary symbols) or with one position changed in exactly one way (if the error is given by a pair of complementary symbols).

Therefore $4^2\,\widehat{t}_n \le \widehat{t}_{n+2} + 4^{m+1}\left[4\binom{m+1}{2} + m + 1\right]$ and, dividing by 4^{n+2}, one obtains $\frac{\widehat{t}_n}{4^n} \le \frac{\widehat{t}_{n+2}}{4^{n+2}} + \frac{4^{m+1}}{4^{n+2}}\left[2m(m+1) + m + 1\right]$.

Referring to the densities $\widehat{\mu}_n$, one has $\widehat{\mu}_{n+2} \ge \widehat{\mu}_n - \frac{1}{4^{m+1}}\left(2m^2+3m+1\right)$. Since $\widehat{\sigma}_n = 1 - \widehat{\mu}_n$, we get $\widehat{\sigma}_{n+2} \le \widehat{\sigma}_n + \frac{1}{4^{m+1}}\left(2m^2+3m+1\right)$. Combining these relations from n to $n+p$, one has

$$\widehat{\sigma}_{n+2p} \le \widehat{\sigma}_n + \sum_{i=m}^{m+p-1} \frac{2i^2+3i+1}{4^{i+1}}.$$

Therefore, the following upper bound for $\widehat{\sigma}$ follows: $\widehat{\sigma} \le \widehat{\sigma}_n + \sum_{i=m}^{\infty} \frac{2i^2+3i+1}{4^{i+1}}$. Hence, for any $n \ge 2$, $n = 2m$, $\widehat{\sigma}_n \le \widehat{\sigma} \le \widehat{\sigma}_n + \sum_{i=m}^{\infty} \frac{2i^2+3i+1}{4^{i+1}}$. The sum can be evaluated by using classical formulas as in the proof of Theorem 13

$$\sum_{i=m}^{\infty} \frac{2i^2+3i+1}{4^{i+1}} = \frac{18m^2+39m+28}{27\cdot 4^m}. \qquad \square$$

4 Comparing Ham- and Lee- isometric quaternary words densities

Let us now compare the density of Ham- and Lee- isometric words in the unique significant case, that is when the alphabet has cardinality $k = 4$. Hence, in this section the value of k will be understood. It turns out that there are more Lee- isometric words than Ham-isometric words. Observe that the result is not obvious. In fact, the set of all Ham-isometric words is not inclusion-wise comparable with the set of Lee-isometric words. Examples are given in [3, 5]. The relation between such sets is described in the next proposition.

Denote $\mathcal{H}_n$ the set of quaternary words of length n having a 2-Ham-error overlap and $\mathcal{L}_n$ the corresponding set for the Lee distance case.

Proposition 19. *Let $f \in \Delta^n$. Then*

- *$f \in \mathcal{H}_n \cap \mathcal{L}_n$ iff f has both a 2-Ham-error overlap and a 2-Lee-error overlap*
- *$f \in \mathcal{H}_n \setminus \mathcal{L}_n$ iff f has a 2-Ham-error overlap and every 2-Ham-error overlap involves pairs of non-complementary symbols, only*
- *$f \in \mathcal{L}_n \setminus \mathcal{H}_n$ iff f has a 2-Lee-error overlap and every 2-Lee-error overlap has only one error position that involves a pair of complementary symbols.*

Proposition 20. *Let $n \ge 2$ be an integer. Then*

- *if $d = 1$ then $\widehat{h_n(d)} = 4^{n-1}$ and $h_n(d) = 0$*
- *If $d \ge 2$ then $\widehat{h_n(d)} = h_n(d) - (5d-7)/2$*

Proof. No 2-Ham-error overlap may have length 1; hence $h_n(1) = 0$. On the other hand, a 2-Lee-error overlap may have length 1. In this case the first and the last symbol in the word are complementary ones. Hence, a word of length n with a 2-Lee-error overlap of length 1 is specified by the first symbol, in 4 ways, and the next $n-2$ symbols. Then, $\widehat{h_n(d)} = 4^{n-1}$. Finally, if $d \ge 2$, the claim follows from Lemmas 8 and 14. $\square$

Unfortunately, as already observed, the previous result cannot be extended to $\widehat{h_n}$ or h_n. The first values of $\widehat{h_n}$ and h_n have been calculated and collected in Table 1. In particular, note that $\widehat{h_n} \leq h_n$, for any $3 \leq n \leq 16$.

Similar calculations show that the density sequences $\widehat{\alpha_n}$ and α_n are not monotonically increasing, already for $n \leq 16$, see Table 1. Let us compare the limit values $\widehat{\sigma}$ and σ. The two following results are consequences of Theorems 13 and 18.

Corollary 21. *The value σ of the limit on the density of Ham-non-isometric quaternary words is*

$$0.833013 \leq \sigma \leq 0.836195$$

Proof. Theorem 13 states that $\sigma_{2m} \leq \sigma \leq \sigma_{2m} + \dfrac{9m^2+15m+8}{18 \cdot 4^{m-2}}$, when $k = 4$. Then, with $m = 8$, it holds that $\sigma_{16} \leq \sigma \leq \sigma_{16} + 0.003182$. Calculations give that $\sigma_{16} = 0.833013$ and finally $0.833013 \leq \sigma \leq 0.833013 + 0.003182 = 0.836195$. □

Corollary 22. *The limit value $\widehat{\sigma}$ of the density of Lee-non-isometric quaternary words is*

$$0.705357 \leq \widehat{\sigma} \leq 0.706200$$

Proof. Taking $m = 8$, $n = 16$, the formula of previous theorem becomes $\widehat{\sigma}_{16} \leq \widehat{\sigma} \leq \widehat{\sigma}_{16} + 0.000843$. Adapting efficient algorithms as in [2, 4], it can be obtained that $\widehat{\sigma}_{16} = 0.705357$ and then $0.705357 \leq \widehat{\sigma} \leq 0.705357 + 0.000843 = 0.706200$. □

The two previous results together allow to compare the limit values σ and $\widehat{\sigma}$ of the densities of Ham- and Lee-non-isometric quaternary words.

Proposition 23. $0.705357 \leq \widehat{\sigma} \leq \sigma \leq 0.836195$.

Let us conclude that the Lee-isometric quaternary words are considerably more than the Ham-isometric ones. In fact, for large n, the number of Ham-isometric words is approximately 17% of all words of that length, whereas the corresponding number for Lee-isometric words is about 30%.

5 Conclusions

Isometric words are at the crossroads of several areas of computer science. They were introduced in the framework of hypercubes and then characterized in terms of overlaps with errors in a word. They can also be defined referring to transformations on words that avoid factors. In this paper, we investigated the density of isometric words defined with respect to Hamming and Lee distances, considering alphabets of any cardinality. Clearly, the results can be restated in terms of the other equivalent characterizations. As a future work, it would be worthwhile to carry out a similar study on the density of isometric words also referring to other distances, as the aforementioned distance based on swap and mismatch operations.

References

1. M. Anselmo, G. Castiglione, M. Flores, D. Giammarresi, M. Madonia, and S. Mantaci. Hypercubes and isometric words based on swap and mismatch distance. In *Descriptional Complexity of Formal Systems. DCFS 2023*, volume 13918 of *Lect. Notes Comput. Sci.*, pages 21–35. Springer, 2023.
2. M. Anselmo, G. Castiglione, M. Flores, D. Giammarresi, M. Madonia, and S. Mantaci. Isometric words based on swap and mismatch distance. In *Developments in Language Theory. DLT23*, volume 13911 of *Lect. Notes Comput. Sci.*, pages 23–35. Springer Nature Switzerland, 2023.
3. Marcella Anselmo, Manuela Flores, and Maria Madonia. Quaternary n-cubes and isometric words. In Thierry Lecroq and Svetlana Puzynina, editors, *Combinatorics on Words*, volume 12842 of *Lect. Notes Comput. Sci.*, pages 27–39. Springer International Publishing, 2021.
4. Marcella Anselmo, Manuela Flores, and Maria Madonia. Fun slot machines and transformations of words avoiding factors. In Pierre Fraigniaud and Yushi Uno, editors, *FUN 2022*, volume 226 of *LIPIcs*, pages 4:1–4:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.
5. Marcella Anselmo, Manuela Flores, and Maria Madonia. On k-ary n-cubes and isometric words. *Theor. Comput. Sci.*, 938(6-7):50–64, 2022.
6. Marcella Anselmo, Dora Giammarresi, Maria Madonia, and Carla Selmi. Bad pictures: Some structural properties related to overlaps. In Galina Jirásková and Giovanni Pighizzini, editors, *DCFS 2020*, volume 12442 of *Lect. Notes Comput. Sci.*, pages 13–25. Springer, 2020.
7. Marie-Pierre Béal and Maxime Crochemore. Checking whether a word is Hamming-isometric in linear time. *Theor. Comput. Sci.*, 933(6-7):55–59, 2022.
8. Marie-Pierre Béal, Filippo Mignosi, and Antonio Restivo. Minimal forbidden words and symbolic dynamics. In *STACS 96, 13th Annual Symposium on Theoretical Aspects of Computer Science*, volume 1046 of *Lecture Notes in Computer Science*, pages 555–566, 1996.
9. Giuseppa Castiglione, Sabrina Mantaci, and Antonio Restivo. Some investigations on similarity measures based on absent words. *Fundam. Informaticae*, 171(1-4):97–112, 2020.
10. Panagiotis Charalampopoulos, Maxime Crochemore, Gabriele Fici, Robert Mercas, and Solon P. Pissis. Alignment-free sequence comparison using absent words. *Inf. Comput.*, 262:57–68, 2018.
11. C. Epifanio, A. Gabriele, F. Mignosi, A. Restivo, and M. Sciortino. Languages with mismatches. *Theoretical Computer Science*, 385(1):152–166, 2007.
12. W.J. Hsu. Fibonacci cubes-a new interconnection topology. *IEEE Transactions on Parallel and Distributed Systems*, 4(1):3–12, 1993.
13. Aleksandar Ilić, Sandi Klavžar, and Yoomi Rho. Generalized Fibonacci cubes. *Discrete Math.*, 312(1):2–11, 2012.
14. Sandi Klavžar. Structure of Fibonacci cubes: A survey. *J. Comb. Optim.*, 25(4):505–522, 2013.
15. Sandi Klavžar and Sergey V. Shpectorov. Asymptotic number of isometric generalized Fibonacci cubes. *Eur. J. Comb.*, 33(2):220–226, 2012.
16. Jianxin Wei. The structures of bad words. *Eur. J. Comb.*, 59:204–214, 2017.