Quasiorthogonal Dimension

Paul C. Kainen[1] and Věra Kůrková[2]

[1] Department of Mathematics and Statistics, Georgetown University
Washington, DC, USA 20057
kainen@georgetown.edu
[2] Institute of Computer Science, Czech Academy of Sciences
Pod Vodárenskou věží 2, 18207 Prague, Czech Republic
vera@cs.cas.cz

Abstract. An interval approach to the concept of dimension is presented. The concept of quasiorthogonal dimension is obtained by relaxing exact orthogonality so that angular distances between unit vectors are constrained to a fixed closed symmetric interval about $\pi/2$. An exponential number of such quasiorthogonal vectors exist as the Euclidean dimension increases. Lower bounds on quasiorthogonal dimension are proven using geometry of high-dimensional spaces and a separate argument is given utilizing graph theory. Related notions are reviewed.

1 Introduction

The intuitive concept of dimension has many mathematical formalizations. One version, based on geometry, uses "right angles" (in Greek, "*ortho gonia*"), known since Pythagoras. The minimal number of orthogonal vectors needed to specify an object in a Euclidean space defines its *orthogonal dimension.*

Other formalizations of dimension are based on different aspects of space. For example, a topological definition (*inductive dimension*) emphasizing the recursive character of d-dimensional objects having $d-1$-dimensional boundaries, was proposed by Poincaré, while another topological notion, *covering dimension*, is associated with Lebesgue. A metric-space version of dimension was developed by Hausdorff, Besicovitch, and Mandelbrot; the new concept of *fractal dimension* can take nonintegral values.

This chapter presents a geometric concept of dimension, using an *interval* approach. We define as in [31, 32, 38], the ε-*quasiorthogonal dimension* of $\mathbb{R}^n$,

$$\dim_\varepsilon(n) := \max\{|X| : X \subset S^{n-1}, x \neq y \in X \Rightarrow |x \cdot y| \leq \varepsilon\} \tag{1}$$

to be the maximum number of unit vectors in $\mathbb{R}^n$ with pairwise-dot-products in the interval $[-\varepsilon, \varepsilon]$ or, equivalently, the maximum number of nonzero vectors whose pairwise angles lie in the interval $[\arccos(\varepsilon), \arccos(-\varepsilon)]$ centered at $\pi/2$.

Interval analysis was introduced by Moore in 1962 [51] and replaces real numbers by intervals. Kreinovich contributed substantially to its modern reformulation as *interval computation* (see Kearfott and Kreinovich [35]), and has been the creator and maintainer of the Interval Computation website [37].

Replacing a "crisp" number by a nontrivial (closed) interval has a profound impact on orthogonal dimension. There are exactly n pairwise-orthogonal nonzero vectors in $\mathbb{R}^n$, but for fixed $\varepsilon > 0$, $\dim_\varepsilon(n)$ grows exponentially with n.

Quasiorthogonality has found numerous applications, including word-space models for semantic classification (Hecht-Nielsen [28], Kaski [33]), selection of input parameters for neural networks (Gorban, Tyukin, Prokhorov, and Sofeikov [22]), estimates of covering numbers (Kůrková and Sanguineti [39]), and prediction of consumer financial behavior (Lazarus [40]).

The chapter is organized as follows. In Section 2 quasiorthogonal dimension is defined and its growth is estimated via geometrical properties of high-dimensional spaces. Section 3 presents a graph theory approach and includes some new results. In Section 4, quasiorthogonal vectors in Hamming cubes are examined. Section 5 describes concrete constructions utilizing sparse ternary vectors. The application of quasiorthogonality to context vectors and computational semantics is in Section 6. The final section includes a number of classical and recent generalizations which are related to quasiorthogonality in other domains.

2 Orthogonal and Quasiorthogonal Geometry

Let $\mathbb{R}^n$ denote the n-dimensional Euclidean space, $S^{n-1} := \{h \in \mathbb{R}^n : \|h\| = 1\}$ is the unit sphere in $\mathbb{R}^n$, and $x \cdot y := \sum_{i=1}^n x_i y_i$ is the inner product of $x, y \in \mathbb{R}^n$.

Hecht-Nielsen introduced prior to 1991 (see [21]) the concept of what was later called a *quasiorthogonal set* (Kůrková and Hecht-Nielsen [38]). For $\varepsilon \in [0, 1)$, a subset T of S^{n-1} is an *ε-quasiorthogonal set* if

$$x \neq y \in T \Rightarrow |x \cdot y| \leq \varepsilon.$$

A set of nonzero vectors is ε-quasiorthogonal if and only if the corresponding set of normalized vectors is ε-quasiorthogonal.

Thus, the *ε-quasiorthogonal dimension of* $\mathbb{R}^n$, $\dim_\varepsilon(n)$, is the maximum cardinality of an ε-quasiorthogonal subset of $\mathbb{R}^n$, i.e., We consider the two cases: (i) ε "small" or (ii) ε "large" w.r.t. $\arcsin(1/n) \sim 1/n$.

In the first case (i), when all pairwise angular measurement errors are *small* (strictly less than $\arcsin(1/n)$), it was shown (Kainen [31], Kainen and Kůrková [32]) that quasiorthogonal dimension equals orthogonal dimension n.

To have a quasiorthogonal set with *more than* n members in n-dimensional Euclidean space, some pair of the vectors must be at an angle which deviates from $\pi/2$ by at least $\arcsin(1/n)$. For instance, for $n = 2$, at least one of the measurements must be in error by at least 30 degrees, corresponding to 1/12-th of a circle. Hence, one can trust an estimate of orthogonal dimension made in a fixed finite-dimensional space if the error is small enough; i.e., precise accuracy in orthogonal dimension is achieved when angular error is sufficiently small.

In the second case (ii), assume that $\varepsilon \in (0, 1)$ is fixed and n increases. It was conjectured in [38] and [28] that ε-quasiorthogonal dimension grows exponentially as n increases. We proved the existence of such exponentially large quasiorthogonal sets using geometry of high-dimensional Euclidean spaces [31] and graph theory [32]), giving the same lower bound on the rate of growth.

We will review both of these approaches, starting with the geometric one. Let E be any set and $\mathcal{F}$ any family of subsets of E; $\mathcal{F}$ is a *packing* if its elements are pairwise-disjoint and $\mathcal{F}$ is a *cover* if its union is E.

For real-valued f and g, we write $f(n) \gtrsim g(n)$ and $f(n) \sim g(n)$ to mean

$$\lim_{n\to\infty} f(n)/g(n) \geq 1 \text{ and } \lim_{n\to\infty} f(n)/g(n) = 1.$$

A simple argument for the existence of large quasiorthogonal sets comes from packing *spherical caps* into the surface of S^{n-1}. The caps consist of all points on the sphere within a fixed angular distance from some center point.

More precisely, let $g \in S^{n-1}$ and let $\varepsilon > 0$. Put

$$C(g,\varepsilon) := \{h \in S^{n-1} \,|\, \langle h, g\rangle \geq \varepsilon\}.$$

Then $C(g,\varepsilon)$ is the set of all unit vectors within angular distance $\alpha = \arccos(\varepsilon)$ from g (see Fig. 1), i.e., the α-ball in the angular metric. As $\varepsilon \to 0^+$, $\arccos(\varepsilon)$ approaches $\pi/2$ from below; that is, the cap is nearly a hemisphere.

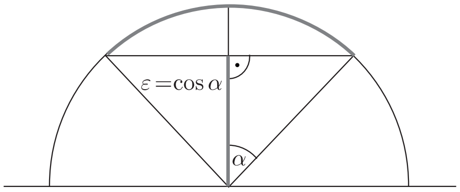

Fig. 1. Spherical cap

Theorem 1. *Let* $0 < \varepsilon < 1$. *Then for all integers* $n \geq 2$,

$$\dim_\varepsilon(n) \geq e^{n\varepsilon^2/2}.$$

Proof. Let μ be the rotationally symmetric uniform probability measure on S^{n-1} obtained by normalizing Lebesgue measure. Determining the area of a cap in Lebesgue measure is well-known (Ball [4, p.11])

$$\mu(C(g,\varepsilon)) \leq \exp\Big(-n\varepsilon^2/2\Big). \tag{2}$$

Hence, any family of such caps which covers S^{n-1} has at least $e^{n\varepsilon^2/2}$ members. Kolmogorov and Tikhomorov [36] showed that the cardinality of a minimum covering by balls of radius r bounds from below the size of a maximum packing by balls of radius $r/2$ but the latter equals $\dim_\varepsilon(n)$ [31, Theorem 2.3]. □

These properties of quasiorthogonality were already implicit in earlier literature on packing spherical caps (Rankin [53] in 1955 and Wyner [63] in 1967) as described in [31] which includes a few other early references not given here.

The upper bound in (2) is quite counter-intuitive since for any fixed ε, the bound becomes very small as n increases. Hence, in high dimension, most of the area of the sphere lies very close to its "equator".

This is a special case of the phenomenon of *concentration of measure*, which states that for large dimensions most of the values of Lipschitz continuous functions concentrate closely around their medians (see, e.g., Matousek [47, p.337]).

Due originally to Levy [42] and Schmidt [60], see also Boucheron, Lugosi, and Massart [7, p. 4], concentration of measure had remained obscure for two decades until it was used by Milman in 1971 to prove a theorem of Dvoretzky which led to the development of the asymptotic theory of normed linear spaces (Milman and Schechtman [49], Ball [4, pp. 41, 47].

Quasiorthogonality is also a special case of the Johnson-Lindenstrauss Lemma [30] on linear projections from spaces of high dimensions to lower-dimensional subspaces that approximately preserve distance on a given finite set.

A function f from $\mathbb{R}^D$ to $\mathbb{R}^d$ is called an *ε-isometry w.r.t. a subset $A \subseteq \mathbb{R}^D$* if for all $a, a' \in A$, f changes square-distances by a multiplicative factor of at most $1 \pm \varepsilon$ - i.e., for all $a, a' \in A$, with $\| \cdot \|$ denoting Euclidean norm,

$$(1-\varepsilon)\|a-a'\|^2 \leq \|f(a)-f(a')\|^2 \leq (1+\varepsilon)\|a-a'\|^2$$

One has the following result from [7, pp. 39–42] .

Lemma 1 (Johnson-Lindenstrauss) *Let A be an n-element subset of $\mathbb{R}^D$ with $\varepsilon, \delta \in (0,1)$. Suppose a random linear mapping $W : \mathbb{R}^D \to \mathbb{R}^d$ is constructed by choosing the dD entries of the standard representing matrix to be normal random variables, centered at zero with variance 1. Then with probability at least $(1-\delta)$, the function W changes the pairwise distances between distinct members of A by a multiplicative factor of at most $1 \pm \varepsilon$ (that is, W is an ε-isometry w.r.t. A)* provided that

$$d \geq \kappa \varepsilon^{-2} \log\left(n\delta^{-1/2}\right)$$

.

The result is essentially sharp and κ is a universal constant which is not larger than 20 [7, p. 41]. Lemma 1 implies that, with high probability, any orthonormal basis of $\mathbb{R}^D$ will be projected to an η-quasiorthogonal set, where $\eta = \varepsilon(2+\varepsilon)$. So in particular $\dim_\eta(d) \geq D$.

A slightly stronger result was given by Dasgupta and Gupta [11], who showed that the following lower bound suffices to guarantee the existence of a linear map W which is an ε-isometry w.r.t. a set of cardinality n.

$$d \geq 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \log(n). \tag{3}$$

They also cited other short proofs of the Johnson-Lindenstrauss Lemma and noted a result [2] of N. Alon showing that (3) is essentially best-possible. Bourgain gave a similar result [9] related to embeddings in Hilbert space. A connection with graphs was exploited by Linial, London, and Rabinovich [44] to obtain bounds on multicommodity flow.

Although the original arguments are nonconstructive, Engebert, Indyk, and O'Donnell [15, Lemma 2] obtain projections by a deterministic algorithm and show that when $S \subset S^{n-1}$, the image of S under the random projection W is an ε-orthogonal set if the projection does not decrease distances in S.

To project, as in Lemma 1, from a high to low-dimensional space, we used a matrix. If the matrix is nearly orthogonal, then distances will be nearly preserved. As quasiorthogonal sets are very common, one can choose the matrix randomly - e.g., with each entry determined by a Gaussian distribution (centered at zero). However, Achlioptas [1] proposed replacing the Gaussian by a discretized distribution taking values $\{-1, 0, 1\}$ with probabilities $(1/6, 2/3, 1/6)$; Bingham and Mannila [6] found that such *sparse random projection* is more efficient, but continues to approximately preserve distance. Li et al. [43] recommend using a much sparser form of Achiloptas' construction where probabilities for each nonzero value are much smaller, and claim a substantial boost to efficiency. For the latest comparisons, see Knoll's thesis [34, p. 46] which considers the similar problem of norm-preservation within a multiplicative factor of $1 \pm \varepsilon$.

3 Graph Theoretic Aspects of Quasiorthogonality

A *graph* G is a symmetric, irreflexive relation (called *adjacency*) on a nonempty set V; equivalently, $G = (V, E)$, where $V = V(G)$ is the set of *vertices* and $E = E(G)$ is the set of *edges*. See, e.g., Harary [27] or Diestel [13] for basic graph theory. For any graph G, a *clique* is a maximal complete subgraph of G and the *clique number* $\omega(G)$ is the largest number of vertices in any clique. If G is connected, then the number of edges in a shortest v-w-path in G defines a *distance* on $V(G)$. The *diameter* $\mathrm{diam}(G)$ of a connected graph G is the greatest distance between any pair of points. The *degree* of a vertex is the number of adjacent vertices. A graph is *r-regular* if all vertices have degree r.

Quasiorthogonality defines a graph by letting adjacency of vertices correspond to quasiorthogonality of vectors. Indeed, let $\emptyset \neq V \subseteq S^{n-1}$ and let $\varepsilon \in [0, 1)$. Define the *ε-orthogonality graph* $G(V, \varepsilon)$ by requiring that

$$\forall v, w \in V = V(G(V, \varepsilon)), \;\; vw \in E(G(V, \varepsilon)) \iff |v \cdot w| \leq \varepsilon.$$

Call H an *orthogonality graph* if H is isomorphic to some $G(V, \varepsilon)$.

What are the basic properties of orthogonality graphs? We provide several such properties below and also show how orthogonality graphs both resemble and differ from random graphs. The first result is in [31] and follows from the strict orthogonality case $\varepsilon = 0$, where it holds by linear algebra - hence the condition on n. Let $\Gamma(n, \varepsilon) := G(S^{n-1}, \varepsilon)$. Then $\omega(\Gamma(n, \varepsilon)) = dim_\varepsilon(n)$.

Theorem 2. *If $n \geq 3$, then $\mathit{\Gamma}(n, \varepsilon)$ has diameter* 2.

If $|V|$ is finite, the diameter-2 condition may not hold. Typically, one would expect the diameter to be quite small [31] but if one chose V to be a finite set of points all very close to a fixed point, then V would induce an edgeless graph.

For any r-regular graph G with p vertices, let

$$\zeta := \zeta(G) := r/(p-1),$$

which is the frequency with which any vertex v is adjacent to the other vertices; ζ is the *density* of the graph. The same notion of density also applies to the orthogonality graph $\mathit{\Gamma}$ with vertex-set S^{n-1} by setting $\zeta(\mathit{\Gamma}) := \mu(W)$, where W is the set of neighbors of v and μ is the probability measure on S^{n-1} obtained by normalizing the Lebesgue measure. By equation (2), we have

Theorem 3. *Let $\varepsilon \in (0,1)$. Then for $\eta := \exp(-n\varepsilon^2/2)$,*

$$\zeta(\mathit{\Gamma}(n,\varepsilon)) \sim 1 - 2\eta. \tag{4}$$

In fact, the same density bound holds for the orthogonality graph induced by the *bipolar* n-vectors from 0 to $\{-1,+1\}^n$; see Theorem 6 below.

Given a positive integer n and $\zeta \in (0,1)$, let $R(n,\zeta)$ denote the *random graph* with n vertices in which the existence of edge vw occurs independently with probability ζ for each distinct pair v, w in V. How close is the n-vertex random graph with probability $\zeta = 1 - 2\eta$ to an orthogonality graph?

By Theorem 2, orthogonality graphs have diameter equal to 2 in many cases and otherwise small. The following result says that a random graph of the same density has very small probability of point-pairs at distance at least 3.

Theorem 4. *Let $v \neq w \in V(R(n,\zeta))$ where $\zeta = 1 - \vartheta$ with $\vartheta \sim 0$. Then*

$$Prob(dist(v,w) \geq 3) = \vartheta(2\vartheta)^{n-2}.$$

Proof. To have distance at least 3 in R, v and w must already be non-adjacent, which has probability ϑ. If u is any vertex other than v and w, then to prevent the existence of a path vuw, not both of $vu \in E$ and $uw \in E$ can hold, which has probability $1 - (1-\vartheta)^2 \sim 2\vartheta$. Further, this must hold for every such vertex u. As edges occur independently, we get the result. □

Thus, orthogonality graphs are rather like dense random graphs in terms of diameter. But orthogonality graphs don't fit the random graph model since adjacency becomes more probable as n increases. One might then expect that with the same number of vertices and the same density, clique number for orthogonality graphs should be larger than for random graphs. However, the next result is in the opposite direction.

Matula [48] proved (in 1976) a very strong clique-size concentration result for random graphs (see also Spencer [61, p. 51]): the clique number is *one of two consecutive integer values*. In the formulation of Bollobas and Erdős [8], the size ω of a maximum clique in any random graph on n vertices with density ζ is

$$\omega(R(n,\zeta)) \sim 2\log(n)/\log(1/\zeta). \tag{5}$$

Theorem 5. *For $\zeta = 1 - 2\eta$ and $\eta = e^{-n\varepsilon^2/2}$,*

$$\omega(R(n,\zeta)) \lesssim \log(n)\dim_\varepsilon(n).$$

Proof. We evaluate $\log(1/\zeta)$. As $\zeta = 1-2\eta$, $1/\zeta \sim 1+2\eta$. But $\log(1+t) \sim t$ for $t \sim 0$. Hence, the denominator in (5) is $\sim 2\eta$. Therefore, one has

$$\omega(R(n,\zeta)) \sim 2\log(n)/2\eta = \log(n)e^{n\varepsilon^2/2} \leq \log(n)\dim_\varepsilon(n);$$

the last inequality is Theorem 1. But $f(n) \sim g(n) \leq h(n) \Rightarrow f(n) \lesssim h(n)$. □

4 Quasiorthogonal Sets in Hamming cubes

Hamming, a founder of information theory, noted that random sets of bipolar vectors (i.e., entries in $\{-1,+1\}$) are almost surely orthogonal [26, p. 188]:

"*For sufficiently large n, there are almost 2^n almost perpendicular lines.*"

Hamming may have meant that the vectors from the origin to the set of all bipolar vectors in n-space form a *probabilistic* clique of size 2^n in the sense that

With probability ~ 1, any pair of bipolar vectors is almost orthogonal.

Hecht-Nielsen and Kůrková, in 1992, conjectured [38] that exponential growth holds for the maximum size of a *strict* clique in which *all* pairs of distinct vectors are approximately orthogonal and introduced the phrase "quasiorthogonal sets".

A proof for exponential growth in the number of pairwise ε-quasiorthogonal vectors in the Hamming cube $H_n := \{-1,1\}^n$, including its rate, was given in 1993 [32]. The argument, sketched after the proof of Theorem 6, uses the Hajnal-Szemeredi Theorem [24] and further guarantees the existence of a large family of such quasiorthogonal sets. Also Theorem 6 follows from Theorem 1.

Recall the notion of *graph complement.* If H is a graph, then the complement $\overline{H}$ is the graph on the same set of vertices as H in which two distinct vertices determine an edge in $\overline{H}$ iff they are *not* adjacent in H, so H and $\overline{H}$ partition the edges of the complete graph on V_H. Under graph complement, cliques correspond to *independent sets* of vertices in which no two vertices are adjacent. Let $\beta(H)$ denote the largest cardinality of any independent set in a graph H so $\beta(H) = \omega(\overline{H})$. Note also that $H = \overline{\overline{H}}$; that is, complement is an involution.

A lower bound on $\beta(H)$ follows from elementary facts about graph coloring as we now show. A *vertex coloring* of a graph H is a partition of its vertices into independent sets. The *chromatic number* $\chi(H)$ of G is the smallest number of parts in such a partition; equivalently, $\chi(H)$ is the least number of "colors" which can be assigned to the vertices of H in such a way that no two vertices of the same color are adjacent.

Recall that for $0 \leq \varepsilon < 1$, let $G(n, \varepsilon)$ and $\mathit{\Gamma}(n, \varepsilon)$ denote the orthogonality graph determined by $V = H_n = \{\pm 1\}^n$ and $V = S^{n-1}$, resp. If $\dim_\varepsilon(n) \sim e^{n\varepsilon^2/2}$, then both inequalities below are asymptotic equalities.

Theorem 6. $\dim_\varepsilon(n) = \omega(\mathit{\Gamma}(n, \varepsilon)) \geq \omega(G(n, \varepsilon)) \gtrsim e^{n\varepsilon^2/2}$.

Proof. (Sketch) The equality is by definition and the first inequality follows from monotonicity of clique number. The second inequality is asymptotic.

For any graph H it is well-known that $\chi(H) \leq 1 + \mathit{\Delta}(H)$, where $\mathit{\Delta}(H)$ denotes the maximum degree of any vertex of H. As the p vertices of H are partitioned into $\chi(H)$ independent sets, at least one of these independent sets has $\geq \lceil p/(1 + \mathit{\Delta}(H)) \rceil$ vertices.

We apply this to the complement of the bipolar orthogonality graph, $H := \overline{G(n, \varepsilon)}$, where independent sets of vertices correspond to quasiorthogonal sets of Hamming vectors. For any two vertices, v and w, there is an isomorphism of H sending v to w so all vertices have the same degree. So we can take $v = (1, 1, \ldots, 1)$ and in the non-orthogonality graph, the degree of v is a sum of binomial coefficients, which can be evaluated by a classical result in information theory (Ash [3, p. 114]) and is $\sim 2^{n\mathcal{H}}$, where $\mathcal{H}$ is the entropy function. Using Taylor's theorem, one gets the result. See [32] for the details. □

Several refinements to this logic can be made.

Let $\beta'(G)$ denote the minimum size of a maximal independent set of a graph G. Clearly, $\beta(G) \geq \beta'(G)$. A theorem of Berge [5, p. 278] states that $\beta'(G) \geq \lceil p/(1+\mathit{\Delta}(G)) \rceil$. So any greedy algorithm which finds a maximal quasiorthogonal set will necessarily produce one of size at least $\lceil p/(1 + \mathit{\Delta}(G)) \rceil$.

In another generalization, Erdős conjectured [5, p. 280] and Hajnal-Szemeredi proved [24] that one can arrange for *each* of the $1 + \mathit{\Delta}(G)$ independent sets in the coloring to have cardinality either $\lceil p/(1 + \mathit{\Delta}(G)) \rceil$ or $\lfloor p/(1 + \mathit{\Delta}(G)) \rfloor$. This is called an *equitable* coloring as color classes differ in size by at most one.

For H_n with $\varepsilon = 1/5$, there are $2^s, s \approx 0.97n$ pairwise-disjoint maximal cliques of size $2^t, t \approx 0.03n$. Is it possible to use this abundance of cliques?

5 Construction of Sparse Ternary Quasiorthognal Sets

In spite of the large number of elements in a quasiorthognal set, one might prefer a specific construction, even of polynomial cardinality, especially if it is an efficient procedure. We will sketch a simple method to achieve this.

A vector is *sparse* if most of its coordinates are zero; we call a vector *ternary* if its entries are -1, 0, and $+1$. The *weight* of a ternary vector is the number of nonzero entries. Sparse ternary vectors are used in studying the co-occurrence of words in models of text semantics. Another application for sparse ternary vectors is in *recommender systems*, where each vector consists, e.g., of a particular user's ratings of movies which are mostly neutral (zero) with a few being $+1$ or -1.

A vector in $\mathbb{R}^n$ is said to have *length* n and is called an n-vector. Given any k-element subset T of $[n] := \{1, \ldots, n\}$ (briefly, k-set in $[n]$), if $2 \leq \ell < k$ is an integer, let $\tau(T, \ell)$ be a maximum size family of ternary n-vectors which are nonzero exactly in the k coordinates in T such that $|v \cdot w| \leq \ell - 1$ if $v \neq w \in \tau(T, \ell)$. Let $t(k, \ell) := |\tau(T, \ell)|$. If $T = \{1, 2, 3\}$, $\ell = 2$ and $n = 6$, then (cf. [31])

$$\tau(T, \ell) = \{(1,1,1,0,0,0), (-,1,-,0,0,0), (1,-,-,0,0,0), (-,-,1,0,0,0)\},$$

where "$-$" denotes "-1".

Start with a maximum family $\mathcal{M}$ of k-sets contained in $[n]$ such that each ℓ-set is in at most one k-set, supposing $2 \leq \ell < k < n$; equivalently, no two members of $\mathcal{M}$ overlap in more than $\ell - 1$ elements. Let $m(n, k, \ell)$ denote the cardinality of $\mathcal{M}$. According to a 1963 conjecture of Erdős and Hanani [16] which was proved by Rődl [55] in 1985, for $k > \ell \geq 2$ fixed, as $n \to \infty$,

$$m(n, k, \ell) \sim \binom{n}{\ell} \Big/ \binom{k}{\ell}. \tag{6}$$

As in [31], let $T(n, k)$ denote the set of all length-n ternary vectors of weight k. Let $T(n, k, \ell)$ be the ε-orthogonality graph with vertex set $T(n, k)$ and $\varepsilon = \frac{\ell-1}{k}$.

Theorem 7. *Let $2 \leq \ell < k$ be integers. For $\varepsilon = k/(\ell - 1)$,*

$$\dim_\varepsilon(n) \geq \omega(T(n, k, \ell)) \geq t(k, \ell) \binom{n}{\ell} \Big/ \binom{k}{\ell}.$$

Proof. As $k^{-1/2}\, T(n, k, \ell) \subset \Gamma(n, \varepsilon)$, the first inequality holds. The second inequality follows from (6). Indeed, for $\mathcal{M}$ as above, put $W := \bigcup_{T \in \mathcal{M}} \tau(T, \ell)$. Then W is a clique in $T(n, k, \ell)$ and has the given number of elements. □

For concreteness, let $\mathcal{F}$ be the family consisting of all 10-sets contained in $[1000]$; $|\mathcal{F}| \approx 2.63 \times 10^{23}$. A subfamily $\mathcal{M}_0$ of $\mathcal{F}$ in which the 10-sets are pairwise disjoint ($\ell = 1$) contains at most 100 elements by the Pigeonhole Principle. But using $k = 10$, $\ell = 3$, according to (6), a maximum subfamily $\mathcal{M}_2 \subseteq \mathcal{F}$ with pairwise overlaps of at most 2 elements has over one *million* elements.

There exists a 12×12 Hadamard matrix, so $t(10, 3) \geq 12$. Replacing each 10-set by $t(10, 3)$ sparse ternary vectors, $\mathcal{M}_2$ generates a clique containing more than 16.6×10^6 vectors whose pairwise normalized dot products do not exceed $1/5$ (hence, the pairwise-angles are between 78 and 102 degrees).

6 Vector Space Models of Word Semantics

The following is a very brief and incomplete account of one of the first scientific areas to utilize quasiorthogonality.

The problem of analyzing word-meaning has taken new significance in the current environment where large amounts of textual information is available online along with powerful computational engines capable of handling a billion-word corpus (Pennington, Socher, and Manning, [52]). A conceptual paradigm, with philosophical roots going back to Wittgenstein, is to group words by their common neighbors. A widely quoted version is *"You shall know a word by the company it keeps,"* due to Firth, a British linguist [18], in 1957.

In order to construct an abstract space, where words can live and in which they can be distributed, vector space ("word space") models with angular distance have been widely used since the SMART (System for the Mechanical Analysis and Retrieval of Text) information retrieval system was developed at Cornell University in the 1960s; see Manning, Raghavan, and Schűtze [46].

Other possibilities could certainly be considered for the analysis of word streams - including graphs, hypergraphs, category-theoretic diagrams, and probabilistic metric-space models - but the vector space approach dominates.

Underlying word-space models is the Distributional Hypothesis (cf. Sahlgren [58]), *Words are similar in meaning if their normalized context vectors are close.*

Context vectors can be formed based on the family of all other words (other than very common and uninformative words such as "and" or "the") or context vectors may utilize multi-word segments (e.g., documents).

If w denotes the number of words and c the number of contexts, then the information structure required is the $w \times c$ *co-occurrence matrix* whose entries can be counts of co-occurrence or normalized frequencies (e.g., how often two words appeared together).

Different techniques can be used to reduce column-dimensionality such as singular value decomposition (SVD), principle components analysis (PCA), or independent component analysis (ICA). However, Sahlgren [57] notes three disadvantages of such techniques: (i) they tend to be computationally infeasible for larger examples, (ii) they need to be repeated each time new data is encountered, and (iii) the initial very-large co-occurrence matrix must still be constructed.

In Random Indexing, one assigns sparse ternary vectors to each context and then the context vectors are summed for each context in which a word appears. This might have significance for classification problems if the nonzero coordinates correspond to some attribute which is either strongly positive or strongly negative. For instance, if the attribute were "connected with animals", then "puppy" would get a $+1$ while "rock" gets -1.

Random projection, as in the Johnson-Lindenstrauss Lemma, has also been used in machine learning and gave results slightly inferior to SVD but with much less effort (Fradkin and Madigan [20] and Li, Hastie, and Church [43]).

7 Some Variants of Orthogonality

The relation of "orthogonality" is important in various fields of mathematics - for example, in combinatorics and functional analysis - not just in geometry.

For n a positive integer, an $n \times n$ array of elements all taken from some fixed n-element set is called a *Latin Square* if each row and each column contains no repeated element. Two order-n Latin Squares A, B are called (LS)*orthogonal* if the ordered superposition

$$\{(A(i,j), B(i,j)) \,|\, i,j = 1, \ldots, n\}$$

contains n^2 distinct elements. Note that A and B may utilize different n-sets for their elements. See, e.g., Dénes and Keedwell [12] and Ryser [56].

Orthogonal Latin Squares were first used for the design of efficient statistical experiments. The largest number of order-n pairwise-orthogonal LS is $n-1$ and, further, the upper bound is achieved when $n \geq 3$ is a prime power; this is also related to the existence of projective planes [56, pp. 79–89].

A notion of "almost orthogonal" LS is described by Mohan in [50] which notes that Horton [29] found two 6×6 Latin Squares whose ordered superposition contains 34 distinct pairs. As Tarry has proved Euler's claim that no pair of order-6 LS is orthogonal, 36 is not achievable. Other ways to weaken orthogonality of LS might also be formulated; see also [12].

Quite different applications of orthogonality and its generalizations occur within analysis. Two measurable functions mapping a measure space (S, μ) to the real numbers are called *orthogonal* if the μ-integral over S of their pairwise-products is zero . As orthogonality implies linear independence, sets of pairwise-orthogonal functions form highly convenient bases for function spaces and are essential to analysis.

Typically, one takes $S = \mathbb{R}^n$ and defines μ by means of a *weighting function.* For example, the vector space of polynomials defined on the real line has, in addition to the usual basis of powers,

$$\{1 = x^0, x, x^2, x^3, \ldots\},$$

a much more useful basis, the Hermite polynomials H_n, which are pairwise-orthogonal with respect to the Gaussian function; that is, for $a \neq b \in \mathbb{N}_+$,

$$\int_{-\infty}^{\infty} H_a(x) H_b(x) \exp(-x^2) dx = 0;$$

see, e.g., Lebedev [41, p. 65].

A notion of quasiorthogonality exists in the case of polynomial functions and was introduced by M. Riesz in 1923. Weakening the condition of orthogonality for infinite sets may still permit partial satisfaction of certain special properties of orthogonal sets of polynomials such as existence of 3-term recursions and locations of zeros. See Brezinsky, Driver, and Redivo-Zaglia [10].

An application of quasiorthogonality in information theory to *space-time block codes* involves concepts simultaneously related to both of the above types of orthogonality; see Farkhani [17] and Su and Xia [62].

A notion of "almost orthogonal" in normed linear spaces is due to Yoshida [65, p. 84], attributed there to F. Riesz in 1918. Let $\|x - A\| := \inf_{a \in A} \|x - a\|$.

Theorem 8. *Let $(X, \|\cdot\|)$ be a normed linear space with $M \neq X$ a closed linear subspace. Then $\forall \varepsilon \in (0,1)$, $\exists x \in X$ with $\|x\| = 1$ and $\|x - M\| \geq 1 - \varepsilon$.*

Yoshida calls x "nearly orthogonal" to M. As a consequence, he gives a short argument for compactness of unit balls in finite dimensional normed linear spaces provided the induced metric is complete in the Cauchy sense, i.e., when $(X, \|\cdot\|)$ is a Banach space.

In a Hilbert space $(X, \cdot)$, with a real inner product, there always exists an orthonormal basis, and every such basis has the same cardinality Schaefer and Wolff [59, p.44], so *vector space dimension* (largest size of a linearly independent set) *equals orthogonal dimension* (largest size of a set of pairwise-orthogonal nonzero vectors). Indeed, pairwise-orthogonal sets of nonzero vectors are independent Deutsch [14, p. 8], while the Gramm-Schmidt orthogonalization procedure [14, pp. 51-52] shows that any linear basis can be converted to an orthonormal basis, so linear and orthogonal dimension coincide.

However, there is a *finite quasiorthogonal dimension* for the Hilbert sphere due to Rankin [54] in 1955. He proved that one can pack only finitely many spherical caps of radius $\rho \in (\pi/4, \pi/2)$ into the set of unit-norm points in Hilbert space; Rankin gives an explicit formula for their number. For a more general approach, applying to Banach spaces, see, e.g., Yan [64]. We conjecture that these packing constants supply bounds on computation which are independent of input dimension.

Following the spherical cap-packing formulation, Zhang [66] uses quasiorthogonality to "develop a fast detection method for a low-rank structure in high-dimensional Gaussian data without using the spectrum information." He bounds *spurious correlation* which occurs when explanatory variables greatly outnumber observations. This situation, where a fixed finite set of data is mapped into increasingly high dimension hypothesis space, is claimed to typically fit a geometric model where data points are vertices of a simplex, which however may be rotated in different ways; see Hall et al. [25]. As a concrete instance, one may have a small number of patient-derived samples which are tested against a large family of genetic hypotheses (Fan et al. [19]).

Acknowledgements

V. Kůrková was partially supported by the Czech Grant Foundation grant GA15-18108S and institutional support of the Institute of Computer Science RVO 67985807. P. C. Kainen received research support from Georgetown University.

References

1. D. Achlioptas, Datbase-friendly random projections, in *ACM Symp. on the Principles of Database Systems*, pp. 274–281, 2001; see also Database-friendly random projections: Johnson-Lindenstrauss with binary coins, *Journal of Comp. & Sys. Sci.*, **66**(4) (2003) 671–687.
2. N. Alon, Problems and results in extremal combinatorics, *Disc. Math* **273**(2003) 1–3.
3. R. B. Ash, *Information Theory*, Dover Publ., New York, 1990 (orig. 1965).
4. K. Ball, An elementary introduction to modern convex geometry, in *Flavors of Geometry*, Ed. by S. Levy, MSRI Publ. 31, Cambridge Univ. Press, 1997, pp. 1–56.
5. C. Berge, *Graphs and Hypergraphs*, North-Holland, Amsterdam, 1973.
6. E. Bingham & H. Mannila, Random projection in dimensionality reduction: Applications to image and text data, in *KDD-2001: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Assoc. Comp. Mach., New York, pp. 245–250.
7. S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A nonasymptotic theory of independence*, Clarendon Press, Oxford, 2012.
8. B. Bollobaś & P. Erdős, Cliques in random graphs, *Math. Proc. Camb. Phil. Soc.* **80** (1976) 419–427.
9. J. Bourgain, On Lipschitz embedding of finite metric spaces in Hilbert space, *Israel J. Math.* **52** (1985) 46–52.
10. C. Brezinsky, K. A. Driver, M. Redivo-Zaglia, Quasi-orthogonality with applications to some families of classical orthogonal polynomials, *Appl. Num. Math.*,**48**(2004)157–168.
11. S. Dasgupta & A.Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss, *Random Struct. & Algor.*, **22**(1) (2003) 60–65.
http://cseweb.ucsd.edu/ dasgupta/papers/jl.pdf
12. J. Dénes & A. D. Keedwell, *Latin Squares and Their Application*, English University Press. London (1974)
13. R. Diestel, *Graph Theory*, volume 173 of *Graduate Texts in Mathematics*, Springer-Verlag, Berlin, third edition, 2005.
14. F. Deutsch, *Best Approximation in Inner Product Spaces*, Springer, New York, 2001.
15. L. Engebretsen, P. Indyk, & R. O'Donnell, Derandomized dimension reduction with applications, *Proc. SODA '02*, (Proc. 13th ann. ACM-SIAM Symp. on Discrete algorithms, San Francisco, 2002), pp. 705–712.
16. P. Erdős & H. Hanani, On a limit theorem in combinatorial analysis, *Publ. Math. Debrecen* **10** (1963), 10–13.
17. J. Farkhani, A quasi-orthogonal space-time block code, *IEEE Trans. on Commun.* **49**(1) (2001) 1–4.
18. J. R. Firth, *Wikipedia*, retrieved 6/9/2017.
19. J. Fan, S. Guo, & N. Hao, Variance estimation using refitted cross-validation in ultrahigh dimensional regression, *J. R. Statist. Soc.* B (2012) **74**, Part 1, 37–65.
20. D. Fradkin & D. Madigan, Experiments with random projections for machine learning, in *Proc. KDD 2003* Proc. 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, Washington, DC, 2003) pp. 517–522.
21. S. I. Gallant, Methods for generating or revising context vectors for a plurality of word stems, *US Patent* US5325298 A, Filing date Sept. 3, 1991; Publ. date June 26, 1994. Assignee: HNC Inc. (Hecht-Nielsen Neurocomputing Corp.)

22. A. N. Gorban, I. Y. Tyukin, D. V. Prokhorov, & K. I. Sofeikov, Approximation with random bases: Pro et contra, *Information Sciences* **364–365** (2016) 129–145.
23. M. Gromov, Isoperimetry of waists and concentration of maps. *GAFA, Geom. Funct. Anal.* **13** (2003) 178–215.
24. A. Hajnal & E.Szemerdi Proof of a conjecture of Erds, In *Combinatorial Theory and Its Applications*, Vol. 2 (Ed. P. Erdős, A. Rényi, & V. T. Sós). North-Holland, Amsterdam, pp. 601–623, 1970.
25. P. Hall, J. S. Marron, & A. Neeman, Geometric representation of high dimension, low sample size data, *J. R. Statist. Soc.*, B **67**(3) (2005) 427–444.
26. R. W. Hamming, *Coding and Information Theory*, Prentice-Hall, Englewood Cliff, NJ, 1986.
27. F. Harary, *Graph Theory* Addison-Wesley, Reading, MA, 1969.
28. R. Hecht-Nielsen, Context vectors: General purpose approximate meaning representations self-organized from raw data, in *Computational Intelligence: Immitating Life*, Eds. J. Zurada, R. Marks & C. Robinson, IEEE Press, 1994, pp. 43–56.
29. J. D. Horton, J.D. (1974). Sub Latin squares and incomplete orthogonal arrays. *J. Comb. Th.* A **16** (1974) 23–33.
30. W. B. Johnson & J. Lindenstrauss, Extensions of Lipschitz maps into a Hilbert space, *Contemp. Math.* **26** (1984), 189–206.
31. P. C. Kainen, Orthogonal dimension and tolerance, *Technical report*, 1992, `https://www.researchgate.net`
32. P. C. Kainen & V. Kůrková, Quasiorthogonal dimension of Euclidean spaces, *Appl. Math. Lett.* **6**(3) (1993) 7–10; `https://www.sciencedirect.com`
33. S. Kaski, Dimensionality reduction by random mapping: Fast similarity computation for clustering, *Proc. 1998 IEEE IJCNN*, 1998. pp. 413–418.
34. F. Knoll, *Johnson-Lindenstrauss Transformations*, Ph.D. Dissertation, Clemson Univ., 2017.
35. R. B. Kearfott & V. Kreinovich, Eds., *Applications of Interval Computations*, Kluwer, Dordrecht, 1996.
36. A. N. Kolmogorov & V. M. Tikhomorov, ε-entropy and ε-capacity of sets in functional spaces, *AMS Transl.* (Ser. 2), **17** (1961) 277–364; orig. *Usp. Mat. Nauk.* **14** (1959)(2) 3–86.
37. V. Kreinovich, *Interval computing*, `http://cs.utep.edu/interval-comp/main.html`
38. V. Kůrková & R. Hecht-Nielsen, Quasiorthogonal sets, *Technical Report* INC-9204, (1992).
39. V. Kůrková & M. Sanguineti, Estimates of covering numbers of convex sets with slowly decaying orthogonal subsets, *Discrete Applied Mathematics* **155** (2007), 1930–1942.
40. M. A. Lazarus et al., Predictive modeling of consumer financial behavior, *US Patent* US6430539 B1, Filing date May 6, 1999; Publ. date Aug. 6, 2002.
41. N. N. Lebedev, *Special Functions and Their Applications*, transl. R. A. Silverman, Dover Publications, Inc., 1972 (orig. Prentice-Hall, 1965).
42. P. Levy, *Problèmes Concrets d'Analyse Functionelle*, Gauthier-Villard, Paris, 1951.
43. P. Li, T. J. Hastie, & K. W. Church, Very sparse random projections, in *Proc. KDD 2006* (Proc. 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 2006), pp.287–296.
44. N. Linial, E. London, & E. Rabinovich, The geometry of graphs and some of its algorithmic applications, *Combinatorica* **15** (2001), 215–246.
45. B. J. MacLennan, Information processing in the dendritic net, pp. 161–197 in *Rethinking Neural Networks: Quantum Fields and Biological Data*, K. H. Pribram, Ed., Lawrence Erlbaum, Hillsdale, 1993.

46. C. D. Manning , P. Raghavan , H. Schűtze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, 2008; online edition, April 1, 2009, https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf
47. J. Matoušek, *Lectures on Discrete Geometry*, Springer, New York, 2002.
48. D. Matula, The Largest Clique Size in a Random Graph, *Technical Report* CS-7608, Dept. of Computer Sci., Southern Methodist Univ., 1976; https://s2.smu.edu/ matula/Tech-Report76.pdf.
49. V. D. Milman & G. Schechtman, *Asymptotic theory of finite dimensional normed spaces*, Lecture Notes in Mathematics 1200, Springer, 1986.
50. R.N.Mohan, M. H. Lee & S. S. Pokhre, On orthogonality of Latin Squares, `arXiv:cs/0604041v2 [cs.DM] (2006).`
51. R.E. Moore, Interval arithmetic and automatic error analysis in digital computing, *Ph.D. Dissertation*, Stanford University, 1962.
52. J. Pennington, R. Socher, & C. D. Manning, GloVe: Global vectors for word representation, *Conference on Empirical Methods in Natural Language Processing* (EMNLP 2014).
53. R. A. Rankin, The closest packing of spherical caps in n dimensions, *Proc. Glasgow Math. Assoc.* **2** (1955) 139–144.
54. R. A. Rankin, On packing of spheres in Hilbert space, *Proc. Glasgow Math. Assoc.* **2** (1955) 145–146.
55. V. Rődl, On a packing and covering problem, *Europ. J. Combinatorics* **5** (1985) 69–78
56. R. J. Ryser, *Combinatorial Mathematics*, Mathematical Assoc. of America, Washington, DC 1963.
57. M. Sahlgren, An introduction to random indexing, in *Methods and Applications of Semantic Indexing Workshop at the 7th Int'l Conf. on Terminology and Knowledge Engineering*, Vol. 87, TermNet News: Newsletter of International Cooperation in Terminology, 2005.
58. M. Sahlgren, The distributional hypothesis, *Rivista di Linguistica* **20**(1) (2008), 33-53.
59. H. H. Schaefer with M. P. Wolff, *Topological Vector Spaces*, Springer, New York, 2nd Edition, 1999.
60. E. Schmidt, E. (1948). Die Brunn-Minkowskische Ungleichung und ihr Spiegelbild sowie die isoperimetrische Eigenschaft der Kugel in der euklidischen und nichteuklidischen Geometrie. *Mathematische Nachrichten*, **1** (1948) 81–115.
61. J. Spencer, *Ten Lectures on the Probabilistic Method*, CBMS-NSF, SIAM, Philadelphia, PA, 1987.
62. W. Su & X-G Xia, Signal constellations for quasi-orthogonal space-time block codes with full diversity, *IEEE Trans. Info. Th.* **50** (10) (2004) 2331–2347.
63. A. D. Wyner, Random packings and coverings of the unit n-sphere, *Bell System Technical J.* **46** (1967) 2111-2118.
64. Y. G. Yan, On the exact value of packing spheres in a class of Orlicz function spaces *J. Convex Analysis* **11**(2) (2004) 394–400.
65. K. Yoshida, *Functional Analysis*, Springer, Berlin, 1965.
66. K. Zhang, Spherical cap packing asymptotics and rank-extreme detection, *IEEE Trans. Info. Th.*, in press 2017; available at https://arxiv.org/pdf/1511.06198.pdf